Predictive and Core-network Efficient RRC Signalling for Active State Handover in RANs with Control/Data Separation

Abdelrahim Mohamed[*], Oluwakayode Onireti[†], Muhammad Imran[†],
Ali Imran[‡], Rahim Tafazolli[*]
[*]Institute for Communications Systems ICS, University of Surrey, Guildford, UK
[†]School of Engineering, University of Glasgow, UK
[‡]School of Electrical and Computer Engineering, University of Oklahoma, USA
E-mail: abdelrahim.mohamed@ surrey.ac.uk

Abstract

Frequent handovers (HOs) in dense small cell deployment scenarios could lead to a dramatic increase in signalling overhead. This suggests a paradigm shift towards a signalling conscious cellular architecture with intelligent mobility management. In this direction, a futuristic radio access network with a logical separation between control and data planes has been proposed in research community. It aims to overcome limitations of the conventional architecture by providing high data rate services under the umbrella of a coverage layer in a dual connection mode. This approach enables signalling efficient HO procedures, since the control plane remains unchanged when the users move within the footprint of the same umbrella. Considering this configuration, we propose a core-network efficient radio resource control (RRC) signalling scheme for active state HO and develop an analytical framework to evaluate its signalling load as a function of network density, user mobility and session characteristics. In addition, we propose an intelligent HO prediction scheme with advance resource preparation in order to minimise the HO signalling latency. Numerical and simulation results show promising gains in terms of reduction in HO latency and signalling load as compared with conventional approaches.

Index Terms

Base stations, cellular networks, context awareness, control data separation architecture, dual connectivity, handover, prediction algorithms, radio access networks.

I. Introduction

Wireless data traffic is increasing dramatically due to proliferation of smart devices and the high dependency on mobile communications in everyday life. Among the possible techniques to overcome the capacity crunch problem, network densification is seen as the most promising solution [1]. As a result, small cells (SCs) are being deployed within the macro cell (MC) coverage to offload some of the users associated with the latter. This is referred to as heterogeneous networks (HetNets) and it is being considered for the long term evolution (LTE) Advanced and beyond. It has been estimated that 50 million base stations (BSs) will be deployed as soon as 2020 [2]. Although these estimations are debatable, they give an indication of the situation in the near future. Such massive deployments raise several problems in terms of signalling overhead, mobility management, energy consumption, capital and running costs, planning and scalability. Most of these issues are tightly coupled to the radio access network (RAN) architecture which constitutes an integral part of cellular systems.

With ultra-dense SC deployments, mobility management becomes complex because handovers (HOs) will happen frequently even for low mobility users. In the conventional RAN architecture, the HO procedure includes transferring all channels (i.e., control and data) from one BS to another with a significant core-network (CN) signalling load [3]–[5]. To solve this problem, a futuristic RAN architecture with a logical separation between control plane (CP) and data plane (DP) has been proposed in research community. In the control/data separation architecture (CDSA), a few MCs, known as control base stations (CBSs), provide the basic connectivity services. Within the CBS footprint, high data rate services are provided by SCs known as data base stations (DBSs). As shown conceptually in Fig. 1, all user equipment (UE) are anchored to the CBS, while the active UE are associated with both the CBS and the DBS in a dual connection mode [5]–[7]. A comprehensive literature survey of this architecture can be found in our paper [8].

This configuration could offer simple and robust HO procedures because the UE is anchored to a BS with a large coverage area. This in turn alleviates mobility signalling and reduces the associated overhead. However, most of the work in this area provides a qualitative discussion rather than a proper analysis with quantitative results. Xu *et al.* [9] and Liu *et al.* [10] argue that the CDSA does not require changing the signalling channel as long as the UE mobility is within the same CBS. Consequently, this could result into minimising mobility management

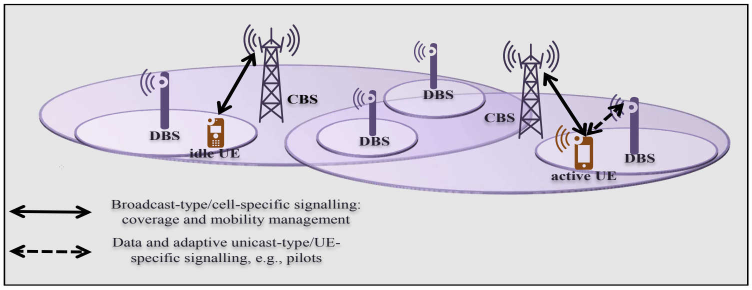

CBS
DBS
idle UE
DBS
DBS
CBS
active UE
DBS
Broadcast-type/cell-specific signalling: coverage and mobility management
Data and adaptive unicast-type/UE-specific signalling, e.g., pilots

Figure 1: Control/Data separation architecture

overhead. Ishii *et al.* [5] and Capone *et al.* [11] adopt a similar approach and discuss potential mobility enhancement opportunities. Zhang *et al.* [12] and the Third Generation Partnership Project (3GPP) studies on dual connectivity [13], [14] depend on simulations to analyse HO failure rate of the CDSA. As a starting point, this paper aims to fill the gap by deriving closed-form expressions for the probability of generating HO-related radio resource control (RRC) CN signalling in both the conventional architecture and the CDSA. This probability can be used to analytically assess the CDSA gains in terms of reduction in the HO-related RRC CN signalling.

From the RAN signalling perspective, the pure CDSA system model may provide marginal gains since a DP HO is always required when the UE moves from one DBS to another. In this direction, context information such as mobility history can play a key role in optimising the RRC and the DBS HO process. It can be used to select the most appropriate DBS for a moving terminal, e.g., a DBS with the highest probability that the user will not leave it quickly [11]. In addition, predicting the DBSs that the UE will visit allows these DBSs to prepare and reserve resources in advance. Such an approach could relax the DBS HO requirements and minimise the associated signalling and interruption time. Nonetheless, such gains might be marginal for non-predictable (i.e., random) users. In fact, a low prediction accuracy could lead to an increase in signalling overhead. Thus, we propose a two mode DBS HO learning and prediction scheme to minimise the HO signalling latency. The following contributions are addressed in this paper:

1) First, we develop an analytical model for the HO-related RRC CN signalling load in both the conventional architecture and the CDSA. Several parameters are incorporated in the model. These include: network deployment parameters, UE contextual information and session characteristics. Markov Chain is utilised to derive the probability of generating RRC CN signalling under general distributions in both architectures. In addition, closed-form expressions are provided in a special case where session duration and cell residence

time are exponentially distributed. This model provides a comparison criteria that can be used to assess the CDSA gains in terms of saving in RRC CN signalling. To the best of our knowledge, this is the first framework that models the HO-related CN efficient RRC signalling in the CDSA.

2) In addition, we propose a mobility prediction model for predictive DBS HO management in CDSA networks. A general learning and prediction scheme that is not restricted to a particular scenario is developed, and we evaluate the signalling cost in both predictive and non-predictive HO management strategies. Moreover, we propose a switching point between predictive and non-predictive schemes based on the prediction entropy.

This paper is an extended version of our work published in [15] where we proposed the DBS HO prediction scheme and assessed its accuracy. In the present paper, we build upon the prediction model in [15] and extend it to include a proactive HO mode selection unit. In addition, we propose the CN efficient RRC signalling scheme and develop an analytical framework to assess the CDSA gains over the conventional architecture. Furthermore, we analyse the HO signalling latency of the integrated solution both analytically and by simulations. The reminder of this paper is structured as follows: Section II discusses the network architecture and introduces a high level overview of the proposed system model. Section III describes the CN efficient RRC signalling scheme, and derives closed-form expressions for RRC CN signalling probability and load. In Section IV, we develop a history-based HO prediction scheme with two modes of operation and formulate its signalling latency cost. Section V presents numerical and simulation results that assess performance of the proposed schemes, while Section VI concludes the paper.

II. Architecture and System Model

A. *Control/Data Separation Architecture*

The main concept of the CDSA depends on separating the signals required for full coverage from those needed to support high data rate transmission. The idea originates from the fact that only small amount of signalling is required to support network connectivity, while data transmission and its related signalling are needed on demand when there are active users. This suggests a two tier RAN architecture with a logical separation between CP and DP, i.e., the CDSA. In the latter, idle UE maintain a single connection with the CBS for network connectivity as shown in Fig. 1. When the UE switches to active mode, e.g., starting a session or receiving a call, it establishes a high rate connection with the DBS whilst maintaining the low rate connection with the CBS for efficient mobility management.

A mapping that illustrates the functionalities supported by each plane can be found in [8, Table II] and [9, Table I]. The CBSs guarantee a low rate coverage layer that provides the necessary signalling related to cell search and acquisition of system information. Broadcast/multicast services, paging functionalities and serving DBS selection are also provided by the CBS. This enables exploiting the large footprint of the CBS and its wider view of network status and parameters which could result into optimised resource selection. During the active session, the DBS provides data transmission along with the necessary signalling for channel estimation, link adaptation and beam-forming. At the same time, the CBS handles mobility management and serves as an RRC anchor point.

B. Proposed Handover Model: High Level Overview

Considering the CDSA described in Section II-A, we propose a signalling efficient mobility management scheme with minimal CN overhead and HO latency. A revisit to the conventional HO scheme is necessary to describe the proposed model. Without loss of generalisation, each conventional HO generates three types of signalling: air interface signalling, RAN signalling and RRC CN signalling. The air interface signalling includes measurement reports that are reported, either periodically or on an event basis, to the serving BS. These reports provide information on signal strength and/or quality of the serving and the neighbouring BSs, based on which HO decisions are made. The UE are informed of these decisions by means of signalling with the serving BS. On the other hand, the HO-related RAN signalling allows the serving and the target BSs to prepare for the HO and exchange the necessary parameters. After accessing the target BS, the data path is switched from the source to the target BS by means of RRC CN signalling. A detailed example for this procedure is provided in Section IV-E.

Fig. 2 shows a high level overview of the proposed HO scheme, while Fig. 3 shows the CDSA layers along with the air interface and backhaul paths in the proposed HO scheme. It utilises the large footprint of the CBS and its functionality as an RRC anchor point in order to minimise the RRC CN signalling load. In this model, the data path from the CN to the RAN remains unchanged as long as the UE mobility is within the same CBS. Such an approach, here referred to as CN efficient RRC signalling, alleviates the HO-related RRC signalling generated towards the

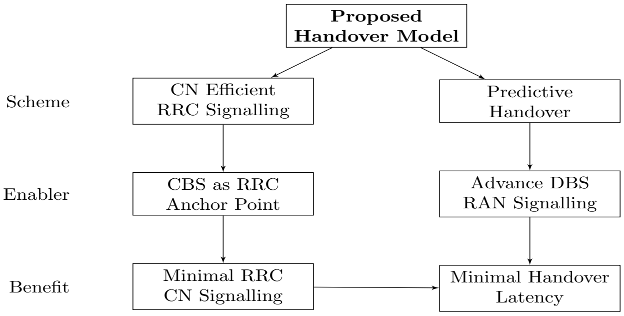


Figure 2: High level overview of the proposed handover model

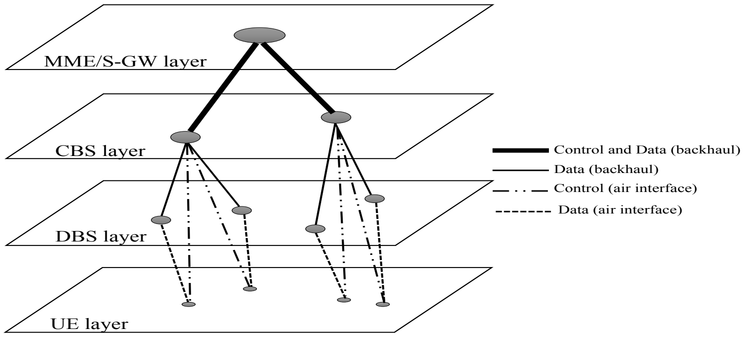


Figure 3: Layers, air interface and backhaul paths in the proposed handover scheme based on control/data separation. Acronym MME: Mobility Management Entity, S-GW: Serving Gateway.

CN since the path switching is performed at the CBS. The analytical modelling of this scheme is provided in Section III. The reduction in RRC CN signalling load minimises the HO overhead and contributes towards reducing the HO latency. The latter is highly dependent on the signalling procedure since each HO signalling message requires some time to be prepared, transmitted and processed at the destination. Thus, we integrate the CN efficient RRC signalling model with a novel HO prediction scheme that enables performing the RAN signalling in advance before the HO criteria is met. Section IV provides a detailed analysis and modelling for the HO prediction scheme and its RAN signalling procedure.

III. Core-network Efficient RRC Signalling Scheme

This section builds upon the CDSA system model and develops the CN efficient RRC signalling scheme. In the CDSA, two types of HOs can be distinguished: intra-CBS HO and inter-CBS HO. The former is the HO between DBSs under the footprint of the same CBS. In other words, the intra-CBS HOs require changing the user-DBS link without changing the user-CBS link. On the other hand, the inter-CBS HO requires changing both the user-DBS link and the user-CBS link, since it is performed between DBSs with different CBS anchor points. As discussed in Section II-B, the HO-related RRC CN signalling is mainly used to switch the DP path when the user performs a HO in the conventional architecture. In the CDSA, however, the CBS is used as an RRC anchor point for the user and as a DP anchor point for the DBSs that are deployed within the CBS footprint. Thus the DP path from the CN to the CBS remains the same as long as the user mobility is within the same CBS, as shown in Fig. 3. Although the intra-CBS HOs require changing the DBS, the DP path from the CBS to the DBS is switched locally at the CBS. As a result, the intra-CBS HOs do not generate RRC CN signalling.

We assume that each HO in the conventional architecture generates RRC CN signalling = C. In the CDSA, each Inter-CBS HO generates RRC CN signalling = S, while intra-CBS HOs do not generate RRC CN signalling as discussed above. Notice that C and S represent the signalling load towards the CN generated by a single HO (in the conventional architecture) and a single inter-CBS HO (in the CDSA), respectively. Thus these quantities may represent the number of the HO-related RRC messages from, to and within the CN, or they can represent the HO-related RRC CN signalling overhead. Here we refer to C and S as RRC CN signalling load irrespective of the measured quantity. The HO-related RRC signalling towards the CN depends on the session duration[1] distribution, UE mobility[2] and BS density (assuming that the transmit power is the same for all cells in the same tier). The expected value of RRC CN signalling load generated by a UE in the CDSA $\mathbb{E}[S]$ can be calculated as:

$$\mathbb{E}[S] = \sum_{i=0}^{\infty} S_i \, f(S_i) \; , \tag{1}$$

where $S_i = i\,S$ and $f(S_i)$ is the probability that the RRC CN signalling load in the CDSA is S_i. This probability can be calculated by using the Markov Chain shown in Fig. 4. Here, P_s^o refers to the probability that the UE will not generate RRC CN signalling in the CDSA, and P_s^g is the probability that the UE will not generate more RRC CN signalling in the CDSA given that it has already generated RRC CN signalling. Since the amount of signalling generated by the UE

[1] It is the time duration between the instance when a session starts and the instance when the session ends, i.e., the total time spent by a UE in active mode for one session.

[2] Here we use the term "cell residence time" to model the UE mobility. It is defined as the total time spent by a UE in a single cell.

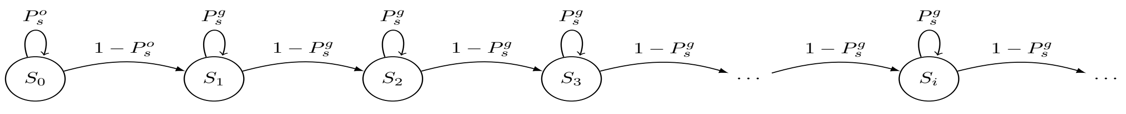


Figure 4: Markov Chain of the handover related RRC core-network signalling

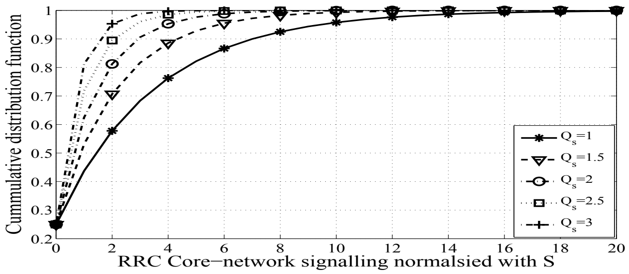


Figure 5: CDF of the RRC CN signalling load in the CDSA, with $P_s^o = 0.25$

increases with the time, a transition from state S_j to state S_i has a zero probability when $j > i$. Based on this model, $f(S_i)$ can be formulated as:

$$f(S_i) = \begin{cases} P_s^o & , \quad \text{for} \quad i = 0 \\ Q_s \, P_s^o \, (1 - P_s^o) \, (1 - Q_s \, P_s^o)^{i-1} & , \quad \text{for} \quad i \geq 1 \end{cases} , \tag{2}$$

where $Q_s \geq 1$ is the ratio between P_s^g and P_s^o. The cumulative distribution function (CDF) of the RRC CN signalling $F(S_i)$ can be written as:

$$F(S_i) = \sum_{j=0}^{i} f(S_j) = P_s^o - \Big((1 - P_s^o) \Big((1 - Q_s \, P_s^o)^i - 1 \Big) \Big) \tag{3}$$

Substituting (2) into (1) and simplifying the resultant equation gives the expected value of RRC CN signalling in the CDSA as a function of P_s^o, Q_s and S:

$$\mathbb{E}\left[S\right] = \frac{S}{Q_s} \left(\frac{1}{P_s^o} - 1 \right) \tag{4}$$

Fig. 5 shows the effect of Q_s on $F(S_i)$. It can be seen that as Q_s increases, the probability of generating large amount of RRC CN signalling decreases while the probability of zero RRC CN signalling load remains constant. In other words, the transition probability from state S_0 to state S_1 increases as Q_s increases, while the transition probability from state S_i to state S_j decreases, where $i \geq 1$ and $j > i$. Thus it can be said that P_s^o and Q_s are important parameters that have a significant influence on the total HO-related RRC CN signalling generated by active UE. Section III-A derives these parameters under general distributions for session duration and cell residence time, while Section III-B provides closed-form expressions in a special case where the session duration and the cell residence time are exponentially distributed.

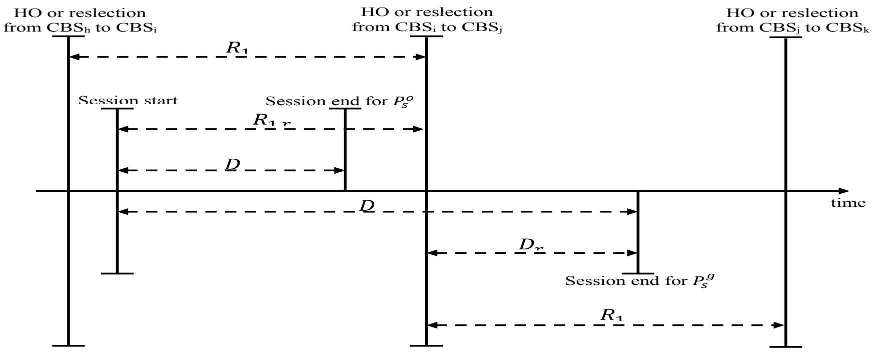

HO or reslection from CBSh to CBSi
HO or reslection from CBSi to CBSj
HO or reslection from CBSj to CBSk
R_1
Session start
Session end for P_s^o
$R_{1,r}$
D
time
D
D_r
Session end for P_s^g
R_1

Figure 6: Timing diagram of the core-network efficient RRC signalling model parameters

A. CDSA with General Distribution for Session Duration and Cell Residence Time

Consider a CDSA cellular network where the CBSs (i.e., the MCs) are modelled as Poisson Point Process (PPP) with density λ_1, while the DBSs (i.e., the SCs) are modelled as another PPP with density λ_2, where $\lambda_2 \geq \lambda_1$. Notice that in typical networks $\lambda_2 \gg \lambda_1$. Assume a session duration D with probability density function (PDF) $f_D(d)$ and mean $\mathbb{E}[D]$. The CBS residence time is modelled as a random variable R_1 with PDF $f_{R_1}(r_1)$ and mean $\mathbb{E}[R_1]$, while the DBS residence time is modelled as a random variable R_2 with PDF $f_{R_2}(r_2)$ and mean $\mathbb{E}[R_2]$. Fig. 6 provides a timing diagram that illustrates the definition of these parameters, without loss of generalisation. We assume that the users move at random directions with a random velocity. Under this assumption, $\mathbb{E}[R_1]$ ($\mathbb{E}[R_2]$) can be approximated by the ratio between the number of UE in a CBS (DBS) and the number of UE leaving a CBS (DBS) per unit of time [16]. Following the derivations in [16], $\mathbb{E}[R_1]$ can be approximated as:

$$\mathbb{E}[R_1] \approx \frac{\pi A_1}{\mathbb{E}[V]\; L_1}\;, \tag{5}$$

where A_1 and L_1 are the average CBS area and perimeter respectively, while $\mathbb{E}[V]$ is the mean velocity. Considering the PPP model, i.e., $A_1 \approx \frac{1}{\lambda_1}$ and $L_1 \approx \frac{4}{\sqrt{\lambda_1}}$ [17], then $\mathbb{E}[R_1]$ can be rewritten as:

$$\mathbb{E}[R_1] \approx \frac{\pi}{4\,\mathbb{E}[V]\,\sqrt{\lambda_1}} \tag{6}$$

Similarly, the mean DBS residence time can be formulated as:

$$\mathbb{E}[R_2] \approx \frac{\pi}{4\,\mathbb{E}[V]\,\sqrt{\lambda_2}} \tag{7}$$

It can be noticed in (5)–(7) that the model requires the mean user velocity rather than the instantaneous value (i.e., irrespective of the velocity distribution). By using the fluid-flow model

[16], different user mobility models can be incorporated in the cell residence time distribution. Consider a UE associated with CBS$_i$ and DBS$_a$. From the CDSA system model in Section II, it can be noticed that the DBS HOs do not generate RRC CN signalling as long as the CBS anchor point remains the same. Thus the definition of P_s^o is equivalent to the probability that the UE does not change CBS$_i$ during the life time of the session. In other words, P_s^o is equivalent to the probability that the session duration is less than the residual residence time in CBS$_i$, i.e.,

$$P_s^o = \text{Prob}\left[D < R_{1,r}\right] = \int\limits_{y=0}^{\infty} f_{R_{1,r}}(y) \int\limits_{x=0}^{y} f_D(x)\,\mathrm{d}x\,\mathrm{d}y \ , \tag{8}$$

where Prob $[\cdot]$ means probability of an event and $R_{1,r}$ is the residual residence time in the CBS as shown in Fig. 6 with PDF $f_{R_{1,r}}(y)$. The latter can be formulated as a function of the CBS residence time distribution based on the residue theorem [18]:

$$\begin{aligned} f_{R_{1,r}}(t) &= \mathcal{L}^{-1}\left\{\frac{1-\mathcal{L}\left\{f_{R_1}(t)\right\}}{s\,\mathbb{E}\left[R_1\right]}\right\} \\ &= \mathcal{L}^{-1}\left\{\frac{4\,\mathbb{E}\left[V\right]\,\sqrt{\lambda_1}\,\left(1-\mathcal{L}\left\{f_{R_1}(t)\right\}\right)}{s\,\pi}\right\} \ , \end{aligned} \tag{9}$$

where $\mathcal{L}$ and $\mathcal{L}^{-1}$ are Laplace transform and inverse Laplace transform operators, respectively.

On the other hand, the definition of P_s^g is equivalent to the probability that the session starts when (or before) the UE is associated with CBS$_i$ and finishes when the UE is associated with CBS$_j$. Thus P_s^g is equivalent to the probability that the residual session duration D_r is less than the CBS residence time, i.e.,

$$P_s^g = \text{Prob}\left[D_r < R_1\right] = \int\limits_{z=0}^{\infty} f_{R_1}(z) \int\limits_{u=0}^{z} f_{D_r}(u)\,\mathrm{d}u\,\mathrm{d}z \ , \tag{10}$$

where $f_{D_r}(u)$ is the PDF of D_r which can be calculated based on the residue theorem as:

$$f_{D_r}(t) = \mathcal{L}^{-1}\left\{\frac{1-\mathcal{L}\left\{f_D(t)\right\}}{s\,\mathbb{E}\left[D\right]}\right\} \tag{11}$$

Q_s can now be calculated as the ratio between (10) and (8). Substituting the resultant Q_s and (8) into (4) gives:

$$\mathbb{E}\left[S\right] = \frac{S\left(1-\int\limits_{y=0}^{\infty} f_{R_{1,r}}(y) \int\limits_{x=0}^{y} f_D(x)\,\mathrm{d}x\,\mathrm{d}y\right)}{\int\limits_{z=0}^{\infty} f_{R_1}(z) \int\limits_{u=0}^{z} f_{D_r}(u)\,\mathrm{d}u\,\mathrm{d}z} \tag{12}$$

B. CDSA with Exponential Distribution for Session Duration and Cell Residence Time

In this section we consider the scenario where the session duration and the cell residence time are exponentially distributed such that

$$f_D(t) = \frac{\mathrm{e}^{-t/\mathbb{E}[D]}}{\mathbb{E}\left[D\right]} \ , \tag{13}$$

and

$$f_{R_1}(t) = \frac{\mathrm{e}^{-t/\mathbb{E}[R_1]}}{\mathbb{E}\left[R_1\right]} = \frac{4\,\mathbb{E}\left[V\right]\sqrt{\lambda_1}}{\pi}\,\mathrm{e}^{-4\,\mathbb{E}[V]\,\sqrt{\lambda_1}\,t/\pi} \tag{14}$$

Lemma 1. *Given that the session duration and the CBS residence time are exponentially distributed, the residual session duration and the residual CBS residence time will also be exponentially distributed.*

Proof. Substituting (13) into (11) and simplifying the resultant equation yields $f_{D_r}(t)$ in the same form as (13). Similarly, substituting (14) into (9) gives $f_{R_{1,r}}(t)$ in the same form as (14). □

The probability that the session duration is less than the residual CBS residence time P_s^o can then be obtained as

$$P_s^o = \frac{\pi}{4\,\mathbb{E}\left[V\right]\,\mathbb{E}\left[D\right]\,\sqrt{\lambda_1} + \pi} \quad , \quad Q_s = 1 \tag{15}$$

by substituting these values into (8) and (10) and solving the integrals. Finally, the expected value of RRC CN signalling load in the CDSA under exponential distribution can be simplified by substituting (13), (14) and the results of Lemma 1 into (12):

$$\mathbb{E}\left[S\right] = \frac{4}{\pi}\,S\,\mathbb{E}\left[V\right]\,\mathbb{E}\left[D\right]\,\sqrt{\lambda_1} \tag{16}$$

C. RRC Core-network Signalling load in Conventional Architecture

The modelling approach proposed in Sections III-A and III-B can be adapted to model the conventional HO signalling in order to assess the CDSA gains. The expected value of RRC CN signalling in the conventional architecture $\mathbb{E}\left[C\right]$ can be written as:

$$\mathbb{E}\left[C\right] = \frac{C}{Q_c}\left(\frac{1}{P_c^o} - 1\right) \ , \tag{17}$$

where P_c^o and $Q_c = \frac{P_c^g}{P_c^o}$ have the same definitions as P_s^o and Q_s respectively, with the subscript c meaning parameters of the conventional architecture. It should be noticed that each HO in the conventional architecture (i.e., MC or SC HO) generates signalling towards the CN because each cell becomes an RRC anchor point. Thus, parameters of the MC layer (i.e., CBS layer in the CDSA terminology) do not capture all the RRC CN signalling load in the conventional architecture. A more convenient design approach is to consider parameters of the SC layer (i.e., DBS layer in the CDSA terminology), since for the very dense deployment considered, the SC HOs $\gg$ MC HOs. As a result, P_c^o, P_c^g and Q_c can be calculated by following a similar approach as the one used in deriving equations (8)$-$(12), by replacing $R_{1,r}$ with $R_{2,r}$, $f_{R_{1,r}}(y)$ with $f_{R_{2,r}}(y)$, $f_{R_1}(y)$ with $f_{R_2}(y)$, and λ_1 with λ_2. The expected value of RRC CN signalling load in the conventional architecture can now be formulated as:

$$\mathbb{E}\left[C\right] = \frac{C\left(1 - \int\limits_{y=0}^{\infty} f_{R_{2,r}}(y) \int\limits_{x=0}^{y} f_D(x)\,\mathrm{d}x\,\mathrm{d}y\right)}{\int\limits_{z=0}^{\infty} f_{R_2}(z) \int\limits_{u=0}^{z} f_{D_r}(u)\,\mathrm{d}u\,\mathrm{d}z} \tag{18}$$

Similarly, when the session duration and the DBS residence time are exponentially distributed, it can be proved that parameters of the conventional architecture simplify to

$$P_c^o = \frac{\pi}{4\,\mathbb{E}\left[V\right]\,\mathbb{E}\left[D\right]\,\sqrt{\lambda_2} + \pi} \quad , \quad Q_c = 1 \tag{19}$$

$$\mathbb{E}\left[C\right] = \frac{4}{\pi}\,C\,\mathbb{E}\left[V\right]\,\mathbb{E}\left[D\right]\,\sqrt{\lambda_2} \tag{20}$$

It can be noticed that the system models of the CDSA and the conventional architecture become memoryless under exponential distribution, since $Q_s = Q_c = 1$ as depicted by (15) and (19). In other words, P_s^g and P_c^g are independent of the previous state and they are equal to P_s^o and P_c^o, respectively. From a signalling load perspective, this can be considered as the worst-case as shown in Fig. 5, thus an appropriate setting of network parameters becomes of great importance in this scenario. The proportional relationship in (20) between $\mathbb{E}\left[C\right]$ and the RRC anchor density in the conventional architecture suggests reducing the latter to minimise the RRC CN signalling load. This can be achieved by moving the RRC anchor point to the CBS of the CDSA in order to exploit the lower density of the CBS since $\lambda_1 \ll \lambda_2$. The CDSA gain G in terms of RRC CN signalling load reduction w.r.t. the conventional architecture can be obtained by:

$$G = 1 - \frac{\mathbb{E}\left[S\right]}{\mathbb{E}\left[C\right]} \tag{21}$$

For the case of $Q_s = Q_c = 1$, the CDSA RRC CN signalling reduction gain can be obtained by substituting (16) and (20) into (21), i.e.,

$$G = 1 - \frac{S}{C}\sqrt{\frac{\lambda_1}{\lambda_2}} \quad , \quad \text{with } Q_s = Q_c = 1 \tag{22}$$

IV. Predictive Handover Model

The CDSA system model may reduce the RRC signalling load towards the CN as discussed in Section III. When complemented with a predictive HO scheme, the CDSA could also minimise the HO related air interface and RAN signalling latency. In this direction, a HO prediction scheme is developed in this section with the main objective of minimising the DBS HO signalling latency.

A. High Level Overview

The HO signalling has been classified in Section II-B into three main components: air interface, RAN and CN signalling. The latter has been tackled in Section III by exploiting the dual connectivity and the centralised CP features of the CDSA. However, the air interface and the RAN signalling remain the same irrespective of the CDSA configuration. Context information and mobility prediction can play a key role in solving these issues by enabling advance signalling for HO preparation and resource reservation. As opposed to the conventional architecture, the CDSA offers relaxed constraints in implementing predictive HO management strategies. These predictive schemes reduce the HO-related air interface signalling by suspending the measurement reports that are transmitted periodically in the conventional architecture. In addition, they reduce the HO latency by enabling the HO-related RAN signalling to be performed in advance before the actual HO criteria is satisfied.

The proposed predictive scheme depends on mobility history to predict future DBS HO events. Fig. 7 shows a block diagram of the DBS HO learning and prediction scheme. Based on a Markov Chain modelling, this scheme uses an online learning process to predict users' trajectory in terms of a DBS HO sequence. The prediction entropy is used as a confidence measure to confirm/reject the predicted DBS. In addition, a recent trajectory dependency parameter is proposed to control the effect of random and less frequent movement patterns. The prediction outcome is utilised to perform the HO-related RAN signalling in advance before the HO criteria is met, resulting into light-weight DBS HO procedures with minimal signalling overhead and latency. Moreover,

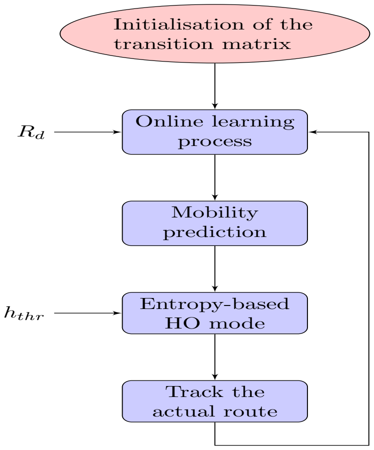


Figure 7: Block diagram of the DBS handover learning and prediction scheme

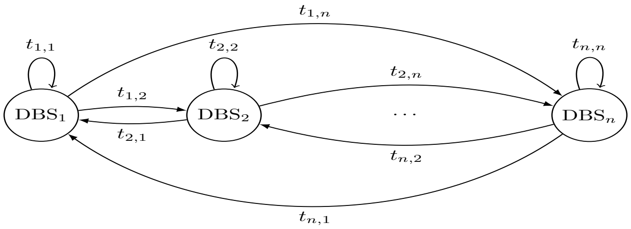


Figure 8: Discrete-time Markov Chain with n states (i.e., DBSs), only states 1, 2 and n are shown for readability

this scheme includes a proactive HO mode selection criteria based on a switching point between predictive HO with advance DBS signalling and non-predictive HO with conventional DBS signalling. This switching point is utilised to detect the unreliable predictions in order to revert back to the conventional signalling mechanism, e.g., when the UE trajectory cannot be reliably predicted or when the UE local HO determination method incorrectly determines successful HOs.

The DP network (i.e., the DBSs) is represented by a discrete-time Markov Chain (DTMC). The latter is a stochastic process characterised by a state space, a transition matrix and an initial distribution [19]. Given the problem under study, a HO from a DBS to another is equivalent to a state transition. Thus each state in the DTMC represents a DBS. Fig. 8 shows a graphical representation of a DTMC with $t_{i,j}$ being the probability of a direct transition (i.e., HO) from DBS$_i$ to DBS$_j$.

The memoryless property of the DTMC implies that the transition matrix would have a static realisation independent of the user's history. In contrast, the proposed model considers a learning transition matrix that can be updated dynamically. Following the derivations of the standard DTMC, the probability distribution can be written as [19]:

$$\mathbf{x}_k = \mathbf{x}_0 \, \mathbf{T}^k \tag{23}$$

with

$$\mathbf{x}_k = [x_1 \; x_2 \; x_3 \; ... \; x_n]$$

$$\mathbf{x}_0 = [\gamma_1 \; \gamma_2 \; \gamma_3 \; \cdots \; \gamma_n]$$

$$\mathbf{T} = \begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,n} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ t_{n,1} & t_{n,2} & \cdots & t_{n,n} \end{bmatrix},$$

where $\mathbf{x}_k$ is the k^{th} HO probability vector, i.e., x_i is the probability of being at DBS$_i$ after k HOs. $\mathbf{T}$ is the transition probability matrix while $\mathbf{x}_0$ is the initial distribution vector with $\gamma_i = 1$ if the user starts the movement at DBS$_i$ and 0 otherwise. Equation (23) can be used to predict a target DBS or a DBS sequence in the user's path. The prediction depends on mobility history which is reflected by $\mathbf{T}$. In the following, we describe the learning procedure for updating the transition matrix.

B. Transition Matrix: Properties and Conditions

Consider $\mathbb{A}$ as the DTMC state space with $\mathbb{I}$ being the states' indices. Define $\mathbb{N}_i$ as a list of the DBSs that are neighbours[3] to DBS$_i \, \forall \, i \in \mathbb{I}$. Notice that $\mathbb{N}_i$ is not a UE-specific parameter, but rather it is a system parameter. The following properties govern $\mathbf{T}$ in the context of the considered transitions (i.e., cellular HOs). These properties are used to set necessary conditions aligned with realistic assumptions.

- Since the number of the DBSs is finite, the DTMC state space is finite:

$$\mathbb{A} = \{\text{DBS}_1, \, \text{DBS}_2, ... \, , \, \text{DBS}_n\} \; , \, \mathbb{I} = \{1, \, 2, ... \, , n\} \tag{24}$$

  where n is the prediction set size. Here we consider DP HO prediction, thus n represents the number of DBSs per CBS.

[3] The first tier neighbours that can be reached directly in a single HO.

- $t_{i,j}$ is a positive real number between 0 and 1 (inclusive):

$$0 \leq t_{i,j} \leq 1 \quad , \quad \forall\, i, j \in \mathbb{I} \tag{25}$$

- A HO from a DBS to itself is not possible. Thus $\mathbf{T}$ is a *hollow matrix*:

$$t_{i,i} = 0 \quad , \quad \forall\, i \in \mathbb{I} \tag{26}$$

- The direct HOs are possible between neighbouring DBSs only:

$$t_{i,j} = t_{j,i} = 0 \quad , \quad \forall\, j \notin \mathbb{N}_i \tag{27}$$

- Any new movement starts from the destination of the previous trajectory. Thus the UE will definitely make an outbound[4] HO from any DBS. However, the UE may not necessarily perform an inbound HO to all the DBSs in the network. As a result, $\mathbf{T}$ is a *right stochastic matrix*. This property sets the following condition:

$$\sum_{j=1}^{n} t_{i,j} = 1 \quad , \quad \forall\, i \in \mathbb{I}. \tag{28}$$

C. Transition Matrix Initialisation

For each user, a $n \times n$ transition matrix is constructed and initialised according to the conditions of Section IV-B. The process of initialising $\mathbf{T}$ involves invoking conditions (26) and (27) to ensure a zero probability for the direct HOs from a DBS to itself or to a non-neighbouring DBS, respectively. Then the remaining elements in $\mathbf{T}$ are initialised with an equi-probable outbound HO assumption, since the new users do not have a mobility history. Algorithm 1 illustrates the initialisation procedure. The initialisation phase is executed only once when the user joins the network for the first time. This phase can be executed either locally by the user or globally by the network.

D. Online Learning Process

The transition matrix can be updated based on each UE DBS HO history. However, maintaining the HO history/frequency for each UE may not be feasible from memory perspective and could lead to an explosion in storage overhead especially in dense deployment scenarios. In order to

[4] The HO from DBS$_i$ to DBS$_j$ is an outbound HO from DBS$_i$ point of view and it is an inbound HO from DBS$_j$ perspective.

Algorithm 1 Initialisation of the transition matrix

```
Invoke conditions (26) and (27).
Set $t_{i,j} = 1$ , $\forall j \in \mathbb{N}_i$.
Set $t_{i,j} = \frac{t_{i,j}}{\sum_{j=1}^{n} t_{i,j}}$ , $\forall i, j \in \mathbb{I}$.
```

reduce the HO latency and improve the performance with minimal storage requirements, we propose an online learning process which gives higher probability to the most common routes. The basic idea is to favour the most common routes followed by the user by giving them higher probabilities compared with other routes. A recent trajectory dependency parameter, R_d, where $0 \leq R_d \leq 1$ is proposed to control the model's reaction to random or less frequent movements. Small (large) values of R_d indicate that the network has a low (high) confidence in the regularity of the user, hence each trajectory will have a low (high) impact on the updated $\mathbf{T}$. The extreme case of $R_d = 0$ means that $\mathbf{T}$ will not be updated (hence the prediction is independent of the movement history), while the case of $R_d = 1$ biases the prediction towards the most recent trajectory.

The process of updating $\mathbf{T}$ can be described by the following example without loss of generalisation. Suppose a user following the path: $\text{DBS}_a \rightarrow \text{DBS}_b \rightarrow \text{DBS}_c$. Then for each HO e.g., from DBS_a to DBS_b, the probabilities of outbound HOs from DBS_a to each neighbouring DBS are updated in a game scheme of several stages. In the first stage, DBS_b and the subset of the DBSs in $\mathbb{N}_a$ that have non-zero probabilities for inbound HOs from DBS_a participate in the game. i.e.,

$$\mathbb{PS}_1 = \left\{ \text{DBS}_j : j \in \mathbb{N}_a \wedge t_{a,j} > 0 \right\} \cup \left\{ \text{DBS}_b \right\}, \tag{29}$$

where $\mathbb{PS}_m$ is the players set in stage $m \geq 1$ of the game. It is worth mentioning that the player set in (29) does not introduce any additional scanning/monitoring load on the UE since this information is already available in current standards. In the LTE for example, the UE periodically measures signal strength and/or signal quality of the serving cell and the top-M other detectable cells at every measurement interval. Thus the player set in (29) can be directly obtained from these measurements.

In the first stage, the probability of the direct HO from DBS_a towards DBS_b is increased by a certain amount controlled by R_d. Similarly, the probabilities of the direct HOs from DBS$_a$ towards all other playing DBSs (i.e., except DBS$_b$) are decreased. This can be expressed mathematically as:

$$t_{a,b}^{(1)} = t_{a,b} + \sum_j t_{a,j}\, R_d \quad , \quad \forall\, \mathrm{DBS}_j \in \mathbb{PS}_1 \setminus \{\mathrm{DBS}_b\} \tag{30}$$

$$t_{a,j}^{(1)} = t_{a,j} - \frac{\sum_j t_{a,j}\, R_d}{|\mathbb{PS}_1| - 1} \quad , \quad \forall\, \mathrm{DBS}_j \in \mathbb{PS}_1 \setminus \{\mathrm{DBS}_b\}\,, \tag{31}$$

where $|\mathbb{PS}_m|$ is the cardinality of the set $\mathbb{PS}_m$, the superscript (m) means the probability after stage m. It can be noticed that the first stage may violate condition (25) because $t_{a,b}$ and $t_{a,j}$ are increased and decreased, respectively, without bounds. A simple solution would be setting a lower bound of 0 and an upper bound of 1 for each entry in $\mathbf{T}$. However this may lead to violating condition (28) because the amount of increase and decrease in the probabilities may not be the same in some cases.

To solve this problem, additional stages are added to reach an equilibrium without violating the conditions of Section IV-B or affecting the learned history. In stage $m > 1$, the DBSs with zero or negative probabilities after stage $m-1$ leave the game. The DBSs with positive probabilities are called *survivals* and they equally share the negative probabilities resulted from stage $m-1$. In other words, the player set in stage $m > 1$ includes the survivals only, i.e.,

$$\mathbb{PS}_m = \left\{ \mathrm{DBS}_j : t_{a,j}^{(m-1)} > 0 \wedge \mathrm{DBS}_j \in \mathbb{PS}_{m-1} \right\}. \tag{32}$$

Since the survivals share the negative entries, their probabilities are equally decreased as:

$$t_{a,j}^{(m)} = t_{a,j}^{(m-1)} + \frac{\sum_n t_{a,n}^{(m-1)}}{|\mathbb{PS}_m|} \ , \ \forall\, \mathrm{DBS}_j \in \mathbb{PS}_m, \tag{33}$$

where $m > 1$, DBS$_n \in \mathbb{PS}_{m-1}$ and $t_{a,n}^{(m-1)} < 0$. Notice that the second term of (33) is negative (i.e., the summation in (33) is for the negative probabilities that resulted from stage $m-1$). Several consecutive stages are added until all the entries in $\mathbf{T}$ are not negative.

Once $\mathbf{T}$ is updated (i.e., after the final stage), the user's trajectory can be predicted by using (23). Given a source DBS where the user starts its current movement, a target DBS or a sequence of candidate DBSs in the user's path can be predicted according to the user's history (which is reflected by $\mathbf{T}$). This can be done by invoking (23) with $k = 1,\ 2,\ 3,\ \ldots$ and $\gamma_i = 1$ for the source DBS, and then selecting the DBS with the highest probability in each HO i.e.,

$$k^{\text{th}}\ \text{HO DBS} = \mathrm{DBS}_w\big|_{x_w = \max(\mathbf{x}_k)}. \tag{34}$$

It is worth mentioning that predictive HO schemes may not be suitable for all users. For instance, the low prediction accuracy of users with highly random mobility profiles may result in increasing the HO latency and the associated signalling overhead. This suggests an adaptive prediction scheme where the user switches between predictive and conventional non-predictive HO procedures, with the main objective of minimising the overall signalling load and latency. Thus we propose a HO mode selection scheme where each prediction is accepted or rejected based on the prediction confidence. The latter can be measured by using the prediction entropy which is a measure of uncertainty, where a higher entropy means higher uncertainty while a zero entropy means full confidence [20]. For the predictive HO scheme, the entropy can be considered as a logarithmic measure for the number of target DBSs with significant probability of being visited, i.e.,

$$h(\mathbf{x}_k) = -\sum_{i=1}^{n} x_i \, \log x_i \ , \tag{35}$$

where $h(\mathbf{x}_k)$ is the entropy of the district probability distribution $\mathbf{x}_k$ given by (23). The unit of the entropy is hartley, where one hartley is the information content of an event if the probability of that event occurring is 10%. Given an entropy threshold h_{thr}, the predictive HO procedure is triggered if there is a high confidence in the predicted target DBS, i.e., $h(\mathbf{x}_k) \leq h_{thr}$. On the other hand, the conventional non-predictive HO procedure is triggered if the entropy of the predicted target DBS does not satisfy the confidence threshold, i.e., $h(\mathbf{x}_k) > h_{thr}$.

The conventional history-based prediction schemes use HO tables, mobility traces and databases to obtain the HO probability. Thus they impose memory requirements in addition to the arithmetic operations used to derive the HO probability. Other prediction schemes do not impose memory requirements but they add a significant computation complexity to predict the HO, e.g., based on Grey models and differential equations. In contrast, the proposed learning and prediction scheme, i.e., (30)$-$(35), consists of basic arithmetic operations with marginal storage requirements because it does not depend on mobility traces.

E. Predictive Handover Latency Cost

As discussed earlier, a reliable prediction of the user's trajectory allows the candidate DBS to prepare and reserve resources in advance, which in turn could simplify the HO process and minimise the associated overhead and interruption time. To investigate this claim, we consider the typical LTE X2 HO procedure as a benchmark for the non-predictive HO scenario. In the latter, the UE measures signals of the detectable DBSs[5] and reports the result to the serving (i.e., the source) DBS whenever the HO criteria is met. The HO procedure consists of three major steps: preparation, execution and completion. In the preparation phase, the source DBS determines the target DBS and establishes a connection with it via the X2 interface. Then the target DBS performs an admission control, reserves resources for the UE and some parameters related to the UE security and ciphering are exchanged between the source and the target DBSs. In the execution phase, the UE detaches from the source DBS and accesses the target DBS. Finally, the HO completion phase switches the DP path towards the target DBS [21].

In the predictive HO procedure, most of the HO preparation steps can be completed before the HO criteria is met, provided that the prediction entropy satisfies the confidence requirements. In this case, the predicted DBS can reserve resources for the UE in advance. Similarly, all the necessary parameters can be exchanged between the source and the predicted DBSs before the HO criteria is met (i.e., advance HO preparation). When the UE sends the measurement report indicating that a HO is required, the source DBS evaluates this report. If the target DBS reported by the UE is the same as the predicted DBS (i.e., correct prediction), then the HO process proceeds with the execution phase. If the prediction is incorrect (i.e., the predicted DBS is not the target DBS being reported by the UE), then the conventional non-predictive HO procedure is triggered. In the latter case, an additional signalling is required to cancel the resources that are reserved in the predicted DBS. Fig. 9 shows the signalling flow diagram for these cases.

The HO signalling cost can be expressed in terms of the delay required to transmit and process the HO messages [22]. Denote $\alpha_{i,j}$ as the one way transmission cost from node i to node j, β_j as the processing cost in node j. The HO signalling latency cost L can be written as [23]:

$$L = \sum \alpha_{i,j} + \sum \beta_j \quad , \tag{36}$$

where the summation in (36) is for all the nodes involved in the signalling flow of the HO process. In other words, each signalling message increases L by a transmission cost $\alpha_{i,j}$ and a processing cost β_j. Since the actual HO procedure starts after the source DBS receives the measurement report, the HO signalling latency cost includes the HO decision and the subsequent

[5] evolved node-B (eNB) in LTE terminology

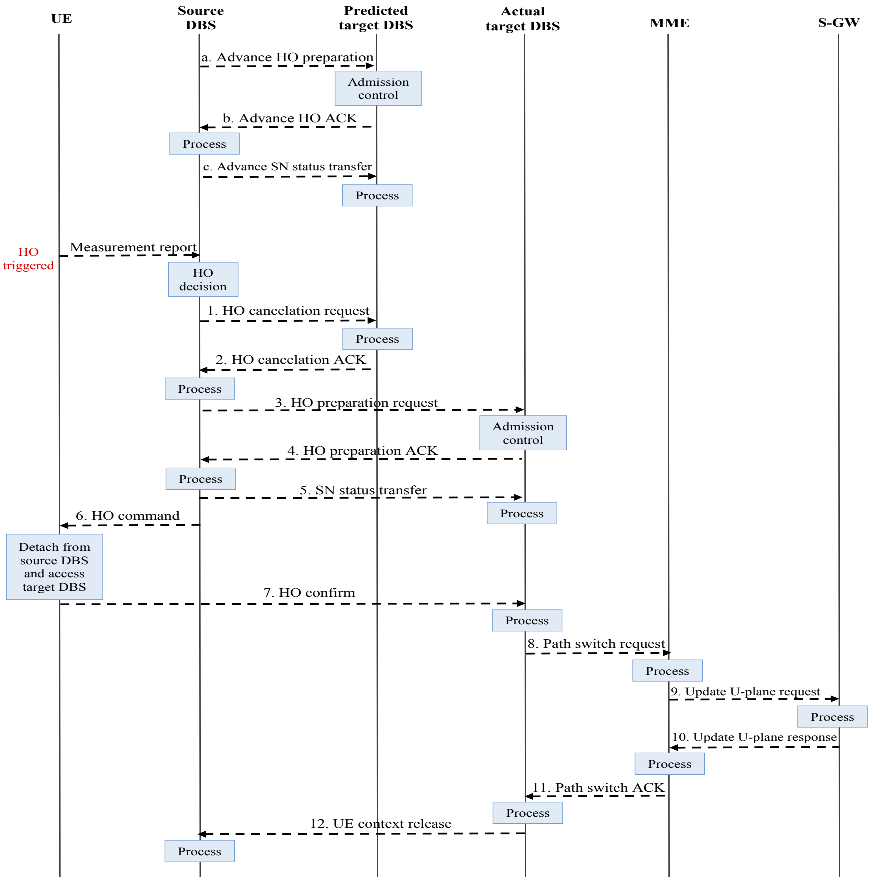


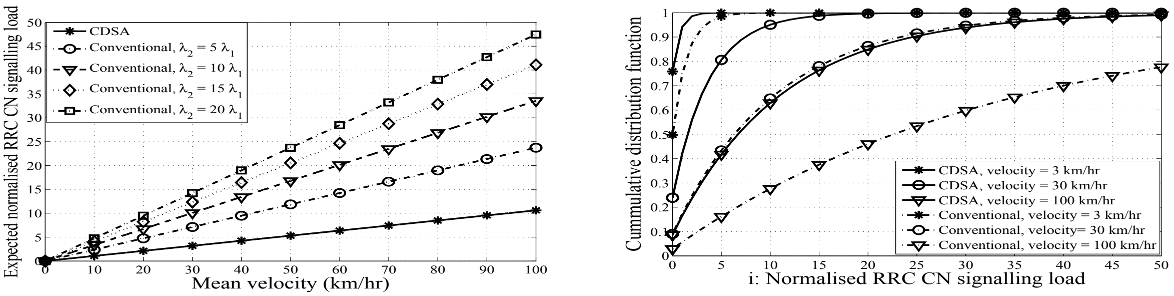
Figure 9: Signalling flow diagram for predictive and non-predictive HO scenarios, based on the LTE X2 HO procedure. Signalling messages in non-predictive HO: 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. Signalling messages in predictive HO with correct prediction: 6, 7, 8, 9, 10, 11, 12. Signalling messages in predictive HO with incorrect prediction: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. For CN efficient RRC intra-CBS HOs: messages 8-11 are not needed and can be replaced with a single ACK message. Acronym ACK: Acknowledgement, SN: Sequence Number.

steps (depending on the HO type). Expressed differently, the advance preparation procedure (i.e., steps a, b and c of Fig. 9) is not included in the cost function of the predictive HO case because the advance reservation phase is completed before the HO is triggered, hence its timing and delay requirements are not critical. The expected cost of the predictive HO L_{pred} can be written as:

$$L_{pred} = A_p \, L_{corr} + (1 - A_p) \, L_{incorr} \ , \tag{37}$$

where A_p is the prediction accuracy, L_{corr} and L_{incorr} are the HO costs with correct and incorrect predictions, respectively, which can be calculated by (36) in conjunction with Fig. 9.

Although the main parameter analysed in this section is the HO signalling latency cost, other



(a) Expected RRC CN signalling load

(b) CDF of RRC CN signalling load with $\lambda_2 = 10\,\lambda_1$

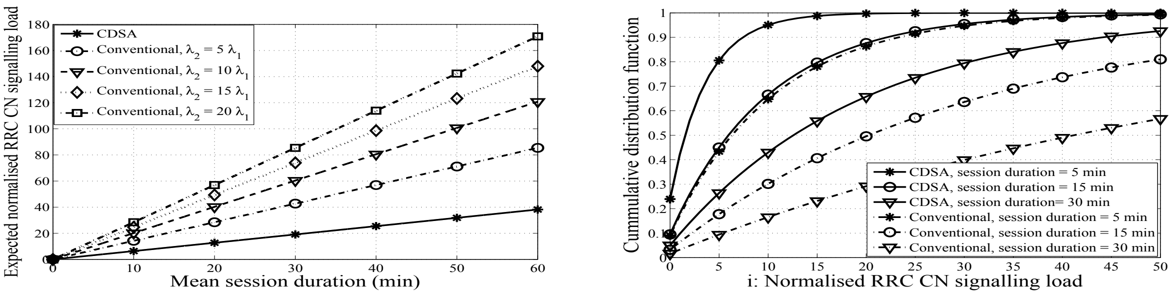
Figure 10: Normalised RRC CN signalling load vs mean velocity, with $\mathbb{E}\left[D\right] = 5$ min and $S = C$

system parameters such as the capacity can be affected by the prediction accuracy. For instance, an incorrect prediction may degrade the overall system capacity, since it reserves resource which could otherwise be used for other users. In addition, the time dimension may have an impact on the overall performance. For example, a too early reservation, even with a correct prediction, wastes the system resources because they are reserved for a long time without being used. However, these aspects are beyond the scope of this paper.

V. Performance Evaluation

A. RRC Core-network Signalling Load Results

This section evaluates the RRC CN signalling of the proposed model in Section III with exponential distribution for session duration and cell residence time. The evaluation is based on normalised densities w.r.t. the CBS density. In addition, the RRC CN signalling load (in terms of expected value and CDF) is normalised with S in the CDSA, and with C in the conventional architecture. Fig. 10a shows the normalised expected value of RRC CN signalling load vs $\mathbb{E}\left[V\right]$ while Fig. 10b provides the CDF of the normalised RRC CN signalling load for low velocity (3 km/hr), medium velocity (30 km/hr) and high velocity (100 km/hr) users, with $\mathbb{E}\left[D\right] = 5$ min, $S = C$ and $\lambda_2 = 10\,\lambda_1$. With a 90% probability, the RRC CN signalling generated in the CDSA with low, medium and high velocity is $\leq \{1,\ 8,\ 25\} \cdot C$ respectively. For the same probability, the RRC CN signalling load in the conventional architecture is $\leq \{3,\ 24,\ 79\} \cdot C$ with low, medium and high velocity respectively. An interesting finding from Fig. 10 is that the RRC CN signalling generated in the CDSA with high velocity (i.e., 100 km/hr) is roughly the same as the signalling generated in the conventional architecture with medium velocity (i.e., 30 km/hr).



(a) Expected RRC CN signalling load

(b) CDF of RRC CN signalling load with $\lambda_2 = 10\,\lambda_1$

Figure 11: Normalised RRC CN signalling load vs mean session duration, with $\mathbb{E}\left[V\right] = 30$ km/hr and $S = C$

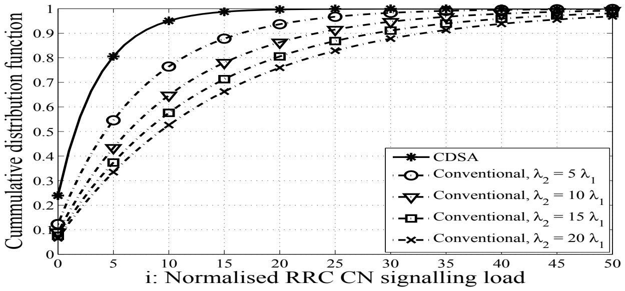


Figure 12: CDF of RRC CN signalling for several DBS densities, with $\mathbb{E}\left[V\right] = 30$ km/hr, $\mathbb{E}\left[D\right] = 5$ min

This can be linked to the CDSA system model and the proportional relationship in (16) between $\mathbb{E}\left[S\right]$ and the term $\mathbb{E}\left[V\right] \cdot \sqrt{\lambda_1}$, and in (20) between $\mathbb{E}\left[C\right]$ and the term $\mathbb{E}\left[V\right] \cdot \sqrt{\lambda_2}$, where $\lambda_1 \ll \lambda_2$. Thus it can be said that the CDSA supports high velocity users with a significantly less RRC CN signalling as compared with the conventional architecture.

Fig. 11a shows the normalised expected value of RRC CN signalling load vs $\mathbb{E}\left[D\right]$ while Fig. 11b provides the CDF of the normalised RRC CN signalling load for several session durations, with $\mathbb{E}\left[V\right] = 30$ km/hr, $S = C$ and $\lambda_2 = 10\,\lambda_1$. As can be seen, the HO related RRC CN signalling in the conventional architecture increases significantly as the session duration increases. Although the CDSA signalling load is also proportional to $\mathbb{E}\left[D\right]$, the latter has a less effect on the CDSA signalling as compared with the conventional architecture.

Fig. 12 shows the effect of the DBS density on the RRC CN signalling load with $\mathbb{E}\left[V\right] = 30$ km/hr, $\mathbb{E}\left[D\right] = 5$ min and $S = C$. It can be noticed that the RRC CN signalling in the conventional architecture increases as λ_2 increases. On the other hand, the CDSA RRC CN signalling load does not depend on the DBS density but rather it depends on the CBS density. With a 90% probability, the CDSA RRC CN signalling is $\leq 8\,C$, while the conventional architecture load is

Table I: Simulation parameters

Parameter	Value
DBS inter-site distance	130 m
DBS transmit power	38 dBm
Transmit mode	SISO (Single Input Single Output)
User density	5 UE/DBS
User speed	10 km/hr for 100% of the users
Measurement gap	200 ms
DBS HO hysteresis	2 dB
Channel model	3GPP Typical Urban [24]
Path loss model	3GPP Urban [25]
Frequency	2 GHz
Bandwidth	10 MHz
Scheduler	Round robin

$\leq \{17,\, 24,\, 29,\, 33\} \cdot C$ with $\lambda_2 = \{5,\, 10,\, 15,\, 20\} \cdot \lambda_1$ respectively.

System level simulations have been performed to validate the proposed modelling approach in Section III and the conclusion of (22). The considered network topology consists of one mobility management entity (MME) and serving gateway (S-GW), 19 omnidirectional DBSs and $1-4$ CBSs. It has been assumed that the UE-CBS link is error free. In addition, a HO between DBSs under the control of different CBSs triggers a CBS HO. The DBS HO criteria follows the signal strength based HO approach, i.e., a DBS HO is triggered if the candidate DBS signal strength is higher than the summation of the serving DBS signal strength and a HO hysteresis. The HO procedure follows the signalling flow without prediction, as illustrated in Fig. 9. Other simulation parameters are provided in Table I.

Fig. 13 compares the theoretical and the simulated CDSA gain in terms of RRC CN signalling load reduction w.r.t. the conventional architecture, while Fig. 14 shows the theoretical gain for other density and configuration values. In the latter, a positive gain means a reduction while a negative gain means an increase in the signalling load. As can be seen in Fig. 13, the gain values obtained from the simulation are in line with the theoretical values which validates the proposed modelling approach. When $S = C$, the CDSA reduces the RRC CN signalling by $55-80\%$ w.r.t. the conventional architecture as shown in Fig. 14. This result indicates that the CDSA is more

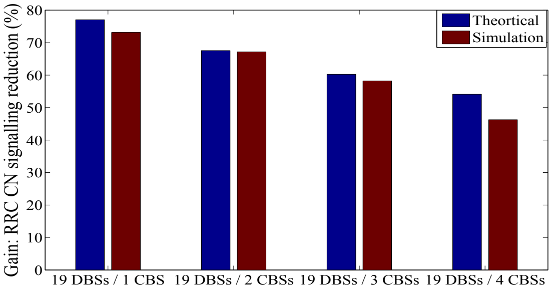


Figure 13: CDSA gain, theoretical vs simulations

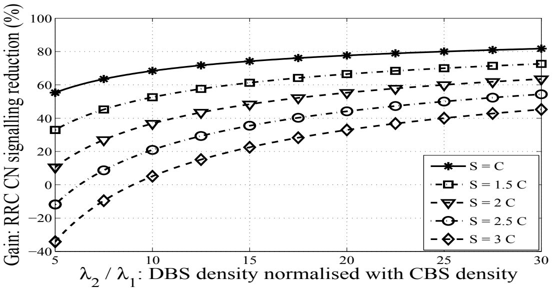


Figure 14: Theoretical CDSA gain vs DBS density

beneficial in dense deployment scenarios as the gain therein is higher.

It is worth mentioning that there is no standard procedure for inter-CBS HO. The later requires changing both the DBS and the CBS, hence the signalling load generated towards the CN by a single inter-CBS HO in the CDSA could be higher than the load generated by a single HO in the conventional architecture. As a result, we evaluate several cases and assume that $S \geq C$. As can be noticed in Fig. 14, the CDSA provides positive gains when the inter-CBS HO procedure generates double the RRC CN signalling load that is generated by a single HO in the conventional architecture. Furthermore, a positive gain can be achieved with $\lambda_2 \geq 9\,\lambda_1$ even if the inter-CBS HO procedure generates 3 times the signalling load generated by the conventional HO procedure. Thus it can be concluded that in dense deployment scenarios the CDSA can significantly reduce the overall RRC signalling towards the CN even if the inter-CBS HO procedure is more complicated than the conventional procedure.

B. Predictive Handover and Latency Results

1) Entropy-based handover mode selection statistics: A second set of system level simulations have been performed to assess performance of the proposed predictive HO scheme. Traces for 100 consecutive days are collected where the trajectory of each day consists of 10 HOs, and the network consists of 69 DBSs under the control of one CBS. We consider a regular user that follows the same route every day (i.e., the HO traces have 0% random mobility), and a user that follows a regular route in some days and random routes in other days. The percentages of the random days w.r.t. the total period are 10%, 20% and 30% and they are distributed evenly across the observation period. The prediction of each day's trajectory is based on the history learned up to the previous day.

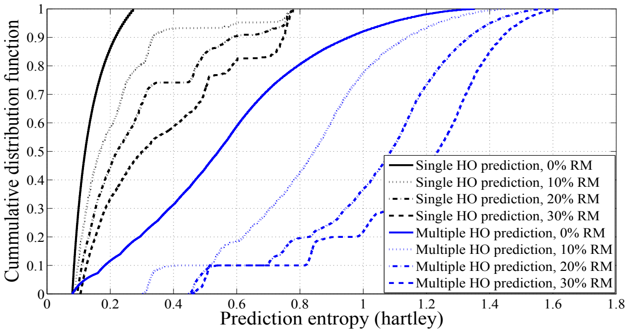
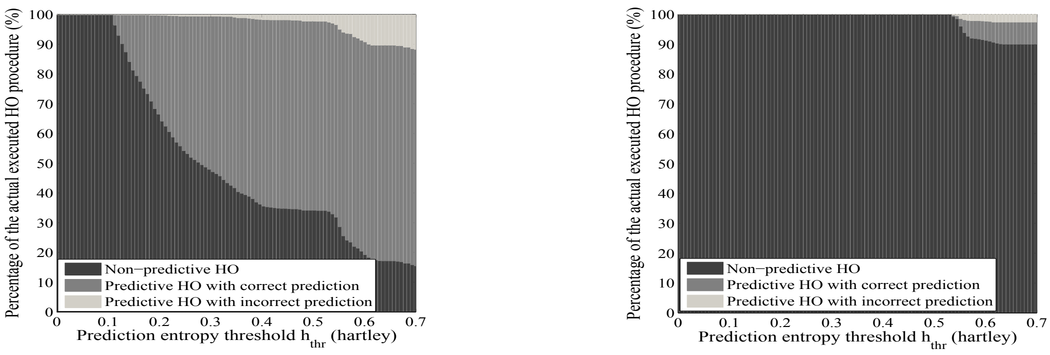


Figure 15: CDF of the prediction entropy. Acronym RM: Random Mobility

Fig. 15 provides the CDF of the prediction entropy for both single and multiple HO predictions. It can be noticed that the entropy of a single HO prediction is significantly less than the entropy of multiple HO prediction. In other words, the prediction of a single HO has a higher confidence than the prediction of multiple HOs. For instance, the prediction entropy of the regular movement scenario (i.e., 0% random mobility) has a 90-th percentile of 0.22 hartley and 0.96 hartley for single and multiple HO prediction, respectively. This can be traced to the fact that the probability of incorrect prediction is higher in the multiple HOs case due to the error propagation and the large number of candidate target DBSs. As opposed to the multiple HO prediction, the single HO prediction does not require matrix multiplication or computationally complex operations. Thus from complexity and performance perspectives, predicting a single HO at a time might be more suitable for practical systems. The random mobility effect can also be seen in Fig. 15, where the 90-th percentile of the prediction entropy increases from 0.22 hartley with 0% random mobility to 0.74 hartley with 30% random mobility, for the single HO prediction case. Expressed differently, the prediction confidence decreases as the UE randomness increases.

The effect of h_{thr} on the switching point between predictive and non-predictive HOs can be seen in Fig. 16 which provides statistics of the actual executed HO type for the 30% random mobility scenario. Considering the single HO prediction case with $h_{thr} = 0.3$ hartley, it can be noticed in Fig. 16a that 50% of the predictions satisfy the predictive HO triggering condition (i.e., $h(\mathbf{x}_k) \leq h_{thr}$). This can be linked to the high confidence (i.e., low entropy) of the single HO prediction depicted by Fig. 15. In this case, the HO preparation phase can be executed in advance for 50% of the HOs, where 96% of them were correct predictions. On the other hand, Fig. 16b shows that all of the multiple HO predictions do not satisfy the triggering condition when $h_{thr} \leq 0.5$ due to the high entropy of this case. As h_{thr} increases (i.e., the confidence



(a) Single handover prediction

(b) Multiple handovers prediction

Figure 16: Statistics of the actual executed handover type vs h_{thr}, with 30% random mobility

Table II: Cost values for handover signalling messages

Cost description	Value
Transmission cost between DBSs over X2	5
Transmission cost between UE and DBS (include processing)	6.5
Transmission cost between DBS and MME	8.5
Processing cost at DBS	4
Processing cost at MME	$5^{\star}$
Processing cost at S-GW	$5^{\star}$
Cost to detach from the source DBS and access the target DBS	12

* Does not include UE context retrieval of 10 ms.

requirement decreases), more predictions satisfy the triggering condition. As a result, more HOs follow the predictive procedure, at the cost of increasing the number of incorrect predictions.

2) Handover latency cost: In the following, we evaluate potential benefits of the proposed CN efficient RRC signalling with predictive DBS HO scheme in terms of signalling latency cost. For simplicity, we follow [23] by assuming that the transmission cost for different messages between the same source-destination pair is the same irrespective of the message size. Similarly, the processing cost for different messages at the same node is constant. In addition, we assume that the MME and the S-GW are located in the same location, thus the transmission delay between these nodes is negligible. Notice that the MME/S-GW transmission delay may be significant in vertical HOs (i.e., between different radio access technologies), however this case is not considered in this paper. Table II provides the cost values which are based on the feasibility study reported in [21] for the intra-LTE X2 HO procedure.

Fig. 17 shows the HO signalling latency reduction in the integrated predictive and CN efficient

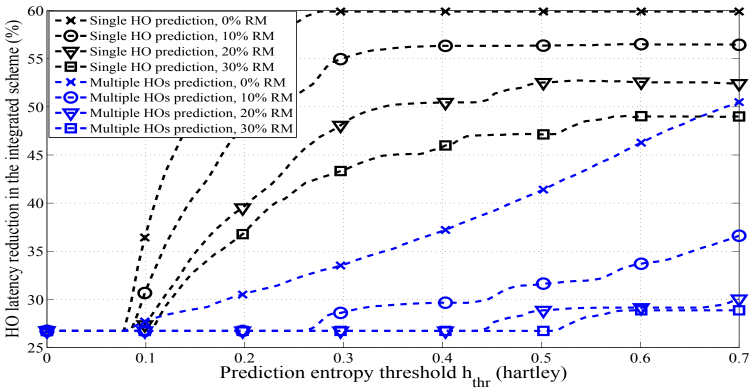


Figure 17: Handover signalling latency reduction in the integrated predictive and CN efficient RRC signalling scheme w.r.t. the conventional handover. Acronym RM: Random Mobility

RRC signalling scheme w.r.t. the conventional HO approach, as a function of the entropy-based switching threshold (i.e., h_{thr}). With a conservative confidence setting of $h_{thr} \leq$ 0.08 hartley, the proposed scheme reduces the HO latency by 26% w.r.t. the conventional HO. This can be linked to Fig. 16 where the high confidence requirement rejects all the predictions, thus all the HOs follow the non-predictive procedure and the gains come from the CN efficient RRC signalling part. Increasing h_{thr} enables more predictions to satisfy the triggering condition, which in turns reduces the signalling latency cost of the integrated scheme by up to 60% for the regular mobility scenario with $h_{thr} \geq$ 0.3 hartley, and by 49% for the 30% random mobility scenario with $h_{thr} \geq$ 0.6 hartley. Thus it can be concluded that h_{thr} is an important design parameter that has a significant effect on the HO signalling latency of the proposed scheme.

VI. Conclusion

In this paper, we developed a signalling efficient HO scheme with minimal overhead and HO latency. The CDSA with dual connectivity is considered as a base architecture, and the CBS-UE link is utilised to move the RRC anchor point to the CBS, resulting into a CN efficient RRC signalling model. An analytical framework was developed to assess superiority of this scheme over the conventional HO approach. Both generic and exponential distributions are considered for session duration and cell residence time, and closed-form expressions are obtained for the expected value of RRC CN signalling load as well as the probability of generating HO-related RRC CN signalling. In addition, a predictive DBS HO scheme with advance RAN signalling was developed in order to minimise the HO latency. It has been found that the proposed integrated scheme, i.e., predictive HO and CN efficient RRC signalling, can significantly reduce the HO-related RRC signalling load and latency. The modelling approach resulted in a signalling load proportional to the velocity and the session duration. Nonetheless, the large CBS footprint (i.e., low CBS density) was found to be an important factor that can reduce the effect of these parameters especially in dense DBS deployment scenarios. Since the predictive HO management strategy is not suitable for users with highly random mobility profiles, the prediction confidence was used as a HO mode decision parameter in order to minimise the overall HO latency.

Acknowledgement

This work was made possible by NPRP grant No. 5-1047-2437 from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors. We would like to acknowledge the support of the University of Surrey 5GIC members for this work.

References

[1] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, “Femtocells: Past, present, and future,” *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, April 2012.

[2] Nokia Siemens Networks, “2020: Beyond 4G, radio evolution for the gigabit experience,” White Paper, August 2011. [Online]. Available: http://nsn.com/file/15036/2020-beyond-4g-radio-evolution-for-the-gigabit-experience

[3] 3GPP, “Study on small cell enhancements for E-UTRA and E-UTRAN: Higher layer aspects,” Technical Report, December 2013, 3GPP TR 36.842 version 12.0.0 Release 12. [Online]. Available: http://www.3gpp.org/DynaReport/36842.htm

[4] ——, “Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility enhancements in heterogeneous networks,” Technical Report, December 2012, 3GPP TR 36.839 version 11.1.0 Release 11. [Online]. Available: http://www.3gpp.org/DynaReport/36839.htm

[5] H. Ishii, Y. Kishiyama, and H. Takahashi, “A novel architecture for LTE-B :C-plane/U-plane split and phantom cell concept,” in *Proc. of IEEE Globecom Workshops*, December 2012, pp. 624–630.

[6] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, “Correlation-based adaptive pilot pattern in control/data separation architecture,” in *Proc. of IEEE International Conference on Communications (ICC)*, June 2015, pp. 2233–2238.

[7] A. Mohamed, O. Onireti, Y. Qi, A. Imran, M. Imran, and R. Tafazolli, “Physical layer frame in signalling-data separation architecture: Overhead and performance evaluation,” in *Proc. of 20th European Wireless Conference*, May 2014, pp. 820–825.

[8] A. Mohamed, O. Onireti, M. Imran, A. Imran, and R. Tafazolli, “Control-data separation architecture for cellular radio access networks: A survey and outlook,” *IEEE Communications Surveys and Tutorials*, vol. 18, no. 1, pp. 446–465, Firstquarter 2016.

[9] X. Xu, G. He, S. Zhang, Y. Chen, and S. Xu, “On functionality separation for green mobile networks: Concept study over LTE,” *IEEE Communications Magazine*, vol. 51, no. 5, pp. 82–90, May 2013.

[10] S. Liu, J. Wu, C. H. Koh, and V. Lau, “A 25 Gb/s(/km2) urban wireless network beyond IMT-advanced,” *IEEE Communications Magazine*, vol. 49, no. 2, pp. 122–129, February 2011.

[11] A. Capone, A. Fonseca dos Santos, I. Filippini, and B. Gloss, “Looking beyond green cellular networks,” in *Proc. of 9th Annual Conference on Wireless On-demand Network Systems and Services*, January 2012, pp. 127–130.

[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. of IEEE Wireless Communications and Networking Conference (WCNC)*, March 2015, pp. 1793–1798.

[13] 3GPP, Nokia Siemens Networks, “Mobility statistics for macro and small cell dual-connectivity cases,” Technical Report, 3GPP TSG-RAN WG2 Meeting , Chicago, USA, 15–19 April 2013. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R2-81b--30048.htm

[14] 3GPP, Huawei, HiSilicon, “Feasible scenarios and benefits of dual connectivity in small cell deployment,” Written Contribution, 3GPP TSG-RAN WG2 Meeting 81, St. Julian’s, Malta, 28 January - 1 February 2013. [Online]. Available: http://www.3gpp.org/DynaReport/TDocExMtg--R2-81--30047.htm

[15] A. Mohamed, O. Onireti, S. Hoseinitabatabae, M. Imran, A. Imran, and R. Tafazolli, “Mobility prediction for handover management in cellular networks with control/data separation,” in *Proc. of IEEE International Conference on Communications (ICC)*, June 2015, pp. 3939–3944.

[16] H. Xie and S. Kuek, “Priority handoff analysis,” in *Proc. of IEEE Vehicular Technology Conference*, May 1993, pp. 855–858.

[17] V. Lucarini, “From symmetry breaking to poisson point process in 2d voronoi tessellations: the generic nature of hexagons,” *Journal of Statistical Physics*, vol. 130, no. 6, pp. 1047–1062, 2008.

[18] Y. Fang, I. Chlamtac, and Y.-B. Lin, “Modeling PCS networks under general call holding time and cell residence time distributions,” *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 893–906, 1997.

[19] J. R. Norris, *Markov chains*. Cambridge university press, 1998.

[20] H. Abu-Ghazaleh and A. Alfa, “Application of mobility prediction in wireless networks using markov renewal theory,” *IEEE Transactions on Vehicular Technology*, vol. 59, no. 2, pp. 788–802, February 2010.

[21] 3GPP, “Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN),” Technical Report, October 2012, 3GPP TR 25.912 version 11.0.0 Release 11. [Online]. Available: http://www.3gpp.org/DynaReport/25912.htm

[22] J. Ho and I. Akyildiz, “Local anchor scheme for reducing signaling costs in personal communications networks,” *IEEE/ACM Transactions on Networking*, vol. 4, no. 5, pp. 709–725, October 1996.

[23] L. Wang, Y. Zhang, and Z. Wei, “Mobility management schemes at radio network layer for LTE femtocells,” in *Proc. of IEEE 69th Vehicular Technology Conference VTC Spring*, April 2009.

[24] 3GPP, “Technical Specification Group Radio Access Network; Deployment aspects,” Technical Report, January 2016, 3GPP TR 25.943 version 13.0.0 Release 13. [Online]. Available: http://www.3gpp.org/DynaReport/25943.htm

[25] ——, “Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios,” Technical Report, January 2016, 3GPP TR 36.942 version 13.0.0 Release 13. [Online]. Available: http://www.3gpp.org/dynareport/36942.htm