Efficient Radio Access Network with Separated Control and Data Functions

Mohamed, Abdelrahim Khalil Aldareer

Submitted for the Degree of Doctor of Philosophy from the University of Surrey

Institute for Communication Systems (ICS)
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH
United Kingdom

December 2016

© Mohamed, Abdelrahim Khalil Aldareer 2016
Abstract

Future cellular systems need to cope with a huge amount of data and diverse service requirements in a flexible, sustainable, green and efficient way with minimal signalling overhead. This calls for network densification, a short length wireless link, efficient and proactive control signalling and the ability to switch off the power consuming devices when they are not in use. In this direction, the conventional always-on service and worst-case design approach has been identified as the main source of inefficiency, and a paradigm shift towards adaptive and on-demand systems is seen as a promising solution. However, the conventional radio access network (RAN) architecture limits the achievable gains due to the tight coupling between network and data access points, which in turn imposes strict coverage and signalling requirements irrespective of the spatio-temporal service demand, channel conditions or mobility profiles. This suggests a new clean slate RAN architecture with a logical separation between the ability to establish availability of the network and the ability to provide functionality or service. This separation of control and data planes provides a framework where limitations and constraints of the conventional RAN can be overcome. In this context, the aim of this thesis is to investigate the control/data separation architecture (CDSA) for futuristic RANs where data services are provided by data base stations (DBSs) under the umbrella of a coverage layer supported by control base stations (CBSs).

A comprehensive literature survey of the CDSA is provided in this thesis. The concept, general structure and basic operation are discussed along with the separation framework and approaches. In addition, limitations of the conventional architecture are pointed out and superiority of the CDSA is discussed whilst focusing on futuristic deployment scenarios. Furthermore, the CDSA technical challenges and enabling technologies are identified, and preliminary standardisation proposals related to this research vision are presented. Three areas, namely energy efficiency, signalling overhead and latency, and mobility management, are identified as promising dimensions that can be substantially improved under CDSA configuration. Focusing on the signalling overhead dimension, a correlation-based adaptive DBS pilot signalling scheme is proposed by exploiting the separation property and the one-to-one nature of the DBS link. The proposed scheme considers channel estimation pilots in the downlink of a multi-carrier DBS air interface, and it depends on estimating the actual channel correlation function to redistribute the pilot signals dynamically. Simulation results show that the proposed adaptive scheme provides a significant saving of 74%–78% in pilot signalling overhead without (or with a marginal) performance penalty as compared with the conventional worst-case design approach.

In addition, the out-of-band signalling related to mobility management is investigated by exploiting the relaxed constraints offered by the CDSA. In particular, the active state handover (HO) signalling in the DBS layer is tackled by proposing two predictive DBS HO signalling schemes with minimal HO latency. These include a history-based predictive DBS HO scheme that predicts future DBS HO events based on the user history, and a measurement-based context-aided predictive DBS HO scheme that predicts future DBS HO events along with the expected HO time by combining DBS signal measurements to physical proximity and user contextual information. In both schemes, the prediction outcome is utilised to perform the HO-related DBS RAN
signalling in advance, resulting into light-weight HO procedures. Simulation results show that these predictive schemes can remarkably reduce the DBS HO signalling latency w.r.t. the benchmark. Precisely, up to 34% reduction in the HO signalling latency is achieved. Moreover, the dual connection feature of the CDSA and the large CBS footprint are utilised to minimise the HO-related core-network (CN) signalling load by proposing a CN-transparent HO signalling scheme. In the latter, the CBS is used as a mobility anchor point for the users and as a data plane anchor point for the DBSs. Thus, the control plane remains unchanged as long as the user mobility is within the same CBS, while the data plane is switched locally at the CBS. Furthermore, the additional data plane backhaul latency induced by the CDSA is modelled and an upper bound for the DBS density under latency constraints is derived. Numerical and simulation results show that the CDSA-based CN-transparent HO signalling scheme significantly outperforms the conventional architecture-based CN-visible HO approaches in terms of CN signalling load. In dense deployment scenarios, the CN-transparent HO scheme is found to be more beneficial where the gains reach 90% reduction in the CN signalling load. Additionally, the CN-transparent HO scheme is integrated with the predictive HO techniques, and simulation results show that the integrated scheme doubles the gains of the predictive-only HO approach in terms of reduction in DBS HO signalling latency.

**Keywords:** Cellular systems; control data separation architecture; mobility management; network densification; radio access network; signalling overhead.

Email:  abdelrahim.mohamed@surrey.ac.uk
Acknowledgements

Thanks God for all the physical and mental strength you bestowed on me during the course of this PhD. I would like to give my sincere gratitude to my supervisors Prof. Muhammad Imran and Prof. Rahim Tafazolli for their support, guidance and sustained effort. Your advices and feedback were crucial in the success of this research. Special thanks to Dr. Atta Quddus for joining the supervisory team and for all the comments and suggestions, and to Prof. Barry Evans for recommending me to this PhD programme and supporting me during my MSc and PhD studies.

Thank you Dr. Oluwakayode Onierti for your valuable comments and contributions. I would also like to thank Dr. Pei Xiao, Prof. Klaus Moessner, Bernard Hunt (Bernie), Dr. Ali Imran, Gerry Foster, Dr. Amir Hoseinitabatabae and the 5G Innovation Centre work area 6 members for their comments, suggestions and discussions surrounding this research. I would like to acknowledge Nicole Steward for her support and help during the course of my PhD.

Last, I owe my deepest gratitude and sincere thanks to my father Khalil Aldareer, my mother Adeela Elmadih, my sisters Enas, Alaa and Fatima and all of my friends for their unlimited encouragement and support.
Declaration of Originality

I hereby declare that the research recorded in this thesis and the thesis itself were composed and originated by myself in the Institute for Communication Systems (ICS), University of Surrey, United Kingdom.

Abdelrahim Mohamed
# Contents

1 Introduction ................................................................. 1  
1.1 Background .............................................................. 1  
1.2 Motivation and Objectives ............................................. 2  
1.3 Overview of Contributions ........................................... 4  
1.4 Thesis Structure and Outline ......................................... 7  
1.5 Publications ............................................................ 9  
1.6 Contributions Outside the Thesis .................................... 10  

2 Background and State-of-the-Art ......................................... 12  
2.1 CDSA Concept and General Structure ............................. 12  
	2.1.1 Basic Operation .................................................... 12  
	2.1.2 CDSA and Software-Defined Networking ...................... 17  
2.2 System Capacity ......................................................... 20  
	2.2.1 Capacity Expansion Mechanisms .............................. 20  
	2.2.2 Heterogeneous Networks ...................................... 22  
	2.2.3 Scalability and Reconfiguration ............................ 23  
2.3 Energy Efficiency ......................................................... 24  
	2.3.1 Conventional Energy Saving Techniques ................... 26  
	2.3.2 Coverage Restrictions ....................................... 27  
	2.3.3 Energy Saving in CDSA ................................ ..... 28  
2.4 Centralised Control Plane .............................................. 30  
	2.4.1 Interference Management ................................... 30  
	2.4.2 Resource Management ........................................ 32  
	2.4.3 Mobility Management .......................................... 33
5 Core-network Transparent Handover Signalling and Backhaul Latency Constraints

5.1 Introduction ......................................................... 100
5.2 System Model ....................................................... 103
5.3 Handover-related Core-network Signalling Load .................... 106
  5.3.1 Core-network Signalling in CN-transparent Handover ........ 106
  5.3.2 Core-network Signalling in Conventional CN-visible Handover 112
5.4 Data Plane Backhaul Latency ...................................... 113
  5.4.1 Data Plane Backhaul Propagation Delay .................... 114
  5.4.2 Data Plane Backhaul Transmission Delay .................... 116
  5.4.3 Data Plane Backhaul Processing Delay .................... 116
  5.4.4 Upper Bound of DBS Density Under Latency Constraints .... 118
5.5 Numerical and Simulation Results .................................. 119
  5.5.1 Core-network Signalling Load Results .................... 119
  5.5.2 DP-BL Constrained DBS Deployment Results ............ 126
  5.5.3 Integrated Handover Signalling Scheme Results ............ 130
5.6 Conclusion ......................................................... 132

6 Epilogue ............................................................... 134

6.1 Thesis Summary and Conclusions .................................. 134
6.2 Future Work ......................................................... 138
  6.2.1 Predictive Mobility Management with DBS Sleep Modes .. 138
  6.2.2 Predictive Mobility Management for CDSA-based Multi-RAT . 139
  6.2.3 Coordinated Adaptive Pilot Signalling Pattern for Multiple Antennas and Neighbouring DBSs ............. 140
  6.2.4 Self-organised Network-driven User Association in CDSA .... 140

Bibliography ............................................................ 142
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Control/Data separation architecture</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Perspectives of the CDSA being studied in international projects</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>Capacity evolution cube</td>
<td>20</td>
</tr>
<tr>
<td>2.4</td>
<td>Typical power consumption of pico BS in LTE system</td>
<td>25</td>
</tr>
<tr>
<td>2.5</td>
<td>Energy and spectral efficiencies vs DBS density</td>
<td>29</td>
</tr>
<tr>
<td>2.6</td>
<td>LTE CRS pattern of four antenna ports</td>
<td>31</td>
</tr>
<tr>
<td>2.7</td>
<td>Conceptual hierarchy of CDSA with RME</td>
<td>33</td>
</tr>
<tr>
<td>2.8</td>
<td>SNR measurements database for channel quality prediction</td>
<td>38</td>
</tr>
<tr>
<td>2.9</td>
<td>Dual connectivity cases</td>
<td>41</td>
</tr>
<tr>
<td>2.10</td>
<td>Transmission/Reception modes to support dual connectivity</td>
<td>44</td>
</tr>
<tr>
<td>3.1</td>
<td>LTE FDD downlink radio frame structure and allocations</td>
<td>48</td>
</tr>
<tr>
<td>3.2</td>
<td>LTE FDD downlink frame overhead, based on 3GPP allocation guidelines</td>
<td>49</td>
</tr>
<tr>
<td>3.3</td>
<td>System model of the proposed adaptive DBS pilot signalling scheme</td>
<td>51</td>
</tr>
<tr>
<td>3.4</td>
<td>Noise effects on the estimated correlation function. SNR = 3 dB</td>
<td>53</td>
</tr>
<tr>
<td>3.5</td>
<td>Example allocations of a DBS downlink resource grid with the proposed adaptive pilot signalling pattern</td>
<td>57</td>
</tr>
<tr>
<td>3.6</td>
<td>Pilot signalling pattern estimation accuracy and potential overhead saving of the proposed adaptive scheme</td>
<td>58</td>
</tr>
<tr>
<td>3.7</td>
<td>BER performance and overhead reduction in the proposed adaptive scheme compared with the LTE CRS pattern</td>
<td>61</td>
</tr>
<tr>
<td>4.1</td>
<td>Signalling flow diagram for predictive and non-predictive HO scenarios</td>
<td>68</td>
</tr>
<tr>
<td>4.2</td>
<td>Discrete-time Markov Chain with ( \hat{N} ) states (i.e., DBSs)</td>
<td>69</td>
</tr>
<tr>
<td>4.3</td>
<td>Block diagram of the history-based DBS handover learning and prediction scheme</td>
<td>72</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>System model of the measurement-based context-aided predictive hand-over scheme</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>Exemplary operation of the signal strength and signal quality prediction unit</td>
<td>78</td>
</tr>
<tr>
<td>4.6</td>
<td>Methods for UE location estimation and CoC/EoC prediction triggering threshold</td>
<td>81</td>
</tr>
<tr>
<td>4.7</td>
<td>Predicted measurement precision with several advance periods, 5 UE per DBS, $V = 10$ km/hr</td>
<td>82</td>
</tr>
<tr>
<td>4.8</td>
<td>Flowchart of the operation flow and decisions of the prediction analysis unit</td>
<td>86</td>
</tr>
<tr>
<td>4.9</td>
<td>Prediction accuracy of the history-based predictive DBS handover scheme</td>
<td>88</td>
</tr>
<tr>
<td>4.10</td>
<td>CDF of the history-based prediction entropy</td>
<td>90</td>
</tr>
<tr>
<td>4.11</td>
<td>Statistics of the actual executed handover type vs $h_{thr}$</td>
<td>91</td>
</tr>
<tr>
<td>4.12</td>
<td>Signalling latency vs $h_{thr}$ in conventional handover and history-based predictive DBS handover</td>
<td>93</td>
</tr>
<tr>
<td>4.13</td>
<td>User speed vs expected handover time referenced to prediction triggering point with a unified triggering threshold for all users</td>
<td>94</td>
</tr>
<tr>
<td>4.14</td>
<td>User speed vs UE-specific prediction triggering threshold</td>
<td>95</td>
</tr>
<tr>
<td>4.15</td>
<td>Prediction statistics of the measurement-based context-aided predictive DBS handover scheme</td>
<td>97</td>
</tr>
<tr>
<td>4.16</td>
<td>Handover signalling latency reduction in the measurement-based context-aided predictive DBS handover scheme w.r.t. the conventional handover</td>
<td>98</td>
</tr>
<tr>
<td>5.1</td>
<td>High level overview of the integrated predictive and core-network transparent handover signalling scheme</td>
<td>104</td>
</tr>
<tr>
<td>5.2</td>
<td>Voronoi spanning trees in CDSA</td>
<td>105</td>
</tr>
<tr>
<td>5.3</td>
<td>Markov Chain of the handover related core-network signalling</td>
<td>107</td>
</tr>
<tr>
<td>5.4</td>
<td>CDF of the normalised CN signalling load in the CN-transparent handover scheme, with $P_1^o = 0.25$</td>
<td>108</td>
</tr>
<tr>
<td>5.5</td>
<td>Timing diagram of the CN-transparent handover model parameters</td>
<td>109</td>
</tr>
<tr>
<td>5.6</td>
<td>Probability of not generating CN signalling vs user speed</td>
<td>120</td>
</tr>
<tr>
<td>5.7</td>
<td>Probability of not generating CN signalling vs session duration</td>
<td>120</td>
</tr>
<tr>
<td>5.8</td>
<td>User speed effect on the normalised CN signalling load</td>
<td>122</td>
</tr>
<tr>
<td>5.9</td>
<td>Session duration effect on the normalised CN signalling load</td>
<td>123</td>
</tr>
<tr>
<td>5.10</td>
<td>DBS density effect on the CDF of the normalised CN signalling load</td>
<td>124</td>
</tr>
</tbody>
</table>
5.11 CN signalling reduction in CN-transparent handover w.r.t. CN-visible handover, theoretical vs simulation .................................................. 124
5.12 Theoretical CN signalling reduction in CN-transparent handover w.r.t. CN-visible handover .................................................. 125
5.13 Effect of CBS and DBS densities on the DP-BL .......................... 127
5.14 Effect of CBS density on the upper bound of DBS density under DB-BL constraints .................................................. 128
5.15 Effect of data load on the upper bound of DBS density under DB-BL constraints .................................................. 129
5.16 Handover signalling latency reduction in the integrated history-based predictive and CN-transparent handover w.r.t. the conventional non-predictive and CN-visible handover .................................................. 131
5.17 Handover signalling latency reduction in the integrated measurement-based context-aided predictive and CN-transparent handover w.r.t. the conventional non-predictive and CN-visible handover .................................................. 131
Acronyms

3GPP    Third Generation Partnership Project
5G      Fifth Generation
ACK     Acknowledgement
AGO     Accumulated Generating Operation
AN      Aggregation Node
AWGN    Additive White Gaussian Noise
BER     Bit Error Rate
BL      Backhaul Latency
BPSK    Bipolar Phase Shift Keying
BS      Base Station
BW      Bandwidth
CBS     Control Base Station
CC      Correlation Coefficient
CDF     Cumulative Distribution Function
CDSA    Control/Data Separation Architecture
CFR     Channel Frequency Response
CN      Core-Network
CoC     Centre of Cell
CoMP    Coordinated Multi-Point
CP      Control Plane
CRS     Cell-specific Reference Signal
CSI-RS  Channel State Information Reference Signal
DBS     Data Base Station
DL      Downlink
DP      Data Plane
DP-BL   Data Plane Backhaul Latency
DRB     Data Radio Bearer
<table>
<thead>
<tr>
<th>Acronyms</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTMC</td>
<td>Discrete-Time Markov Chain</td>
</tr>
<tr>
<td>EC</td>
<td>Energy Consumption</td>
</tr>
<tr>
<td>EE</td>
<td>Energy Efficiency</td>
</tr>
<tr>
<td>eICIC</td>
<td>enhanced Inter-Cell Interference Coordination</td>
</tr>
<tr>
<td>EoC</td>
<td>Edge of Cell</td>
</tr>
<tr>
<td>EPA</td>
<td>Extended Pedestrian-A</td>
</tr>
<tr>
<td>FD</td>
<td>Frequency Domain</td>
</tr>
<tr>
<td>FDD</td>
<td>Frequency Division Duplex</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile communication</td>
</tr>
<tr>
<td>HetNet</td>
<td>Heterogeneous Network</td>
</tr>
<tr>
<td>HO</td>
<td>Handover</td>
</tr>
<tr>
<td>HOF</td>
<td>Handover Failure</td>
</tr>
<tr>
<td>IAGO</td>
<td>Inverse Accumulated Generating Operation</td>
</tr>
<tr>
<td>ICI</td>
<td>Inter-Cell Interference</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>KPI</td>
<td>Key Performance Indicator</td>
</tr>
<tr>
<td>LA</td>
<td>Local Area</td>
</tr>
<tr>
<td>LS</td>
<td>Least Square</td>
</tr>
<tr>
<td>LTE</td>
<td>Long Term Evolution</td>
</tr>
<tr>
<td>M2M</td>
<td>Machine-to-Machine</td>
</tr>
<tr>
<td>MC</td>
<td>Macro Cell</td>
</tr>
<tr>
<td>MIMO</td>
<td>Multiple Input Multiple Output</td>
</tr>
<tr>
<td>mm-wave</td>
<td>Millimetre wave</td>
</tr>
<tr>
<td>MME</td>
<td>Mobility Management Entity</td>
</tr>
<tr>
<td>MR</td>
<td>Measurement Report</td>
</tr>
<tr>
<td>MRC</td>
<td>Measurement Reporting and Control</td>
</tr>
<tr>
<td>NL</td>
<td>Neighbour List</td>
</tr>
<tr>
<td>NMSE</td>
<td>Normalised Mean Square Error</td>
</tr>
<tr>
<td>OFDM</td>
<td>Orthogonal Frequency Division Multiplexing</td>
</tr>
<tr>
<td>OTA</td>
<td>Over-The-Air</td>
</tr>
<tr>
<td>PA</td>
<td>Power Amplifier</td>
</tr>
<tr>
<td>PCC</td>
<td>Phantom Cell Concept</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>PDP</td>
<td>Power Delay Profile</td>
</tr>
<tr>
<td>PL</td>
<td>Physical Layer</td>
</tr>
<tr>
<td>PPP</td>
<td>Poisson Point Process</td>
</tr>
<tr>
<td>PrMR</td>
<td>Predictive Measurement Report</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>QPSK</td>
<td>Quadrature Phase Shift Keying</td>
</tr>
<tr>
<td>RAN</td>
<td>Radio Access Network</td>
</tr>
<tr>
<td>RAT</td>
<td>Radio Access Technology</td>
</tr>
<tr>
<td>RF</td>
<td>Radio Frequency</td>
</tr>
<tr>
<td>RLF</td>
<td>Radio Link Failure</td>
</tr>
<tr>
<td>RME</td>
<td>Resource Management Entity</td>
</tr>
<tr>
<td>RRC</td>
<td>Radio Resource Control</td>
</tr>
<tr>
<td>Rx</td>
<td>Receiver</td>
</tr>
<tr>
<td>S-GW</td>
<td>Serving Gateway</td>
</tr>
<tr>
<td>SC</td>
<td>Small Cell</td>
</tr>
<tr>
<td>SDN</td>
<td>Software-Defined Networking</td>
</tr>
<tr>
<td>SE</td>
<td>Spectral Efficiency</td>
</tr>
<tr>
<td>SINR</td>
<td>Signal-to-Interference-plus-Noise Ratio</td>
</tr>
<tr>
<td>SISO</td>
<td>Single Input Single Output</td>
</tr>
<tr>
<td>SN</td>
<td>Sequence Number</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SON</td>
<td>Self-Organising Network</td>
</tr>
<tr>
<td>SQ</td>
<td>Signal Quality</td>
</tr>
<tr>
<td>SRB</td>
<td>Signalling Radio Bearer</td>
</tr>
<tr>
<td>SS</td>
<td>Signal Strength</td>
</tr>
<tr>
<td>TD</td>
<td>Time Domain</td>
</tr>
<tr>
<td>TDD</td>
<td>Time Division Duplex</td>
</tr>
<tr>
<td>TDM</td>
<td>Time Division Multiplexing</td>
</tr>
<tr>
<td>Tx</td>
<td>Transmitter</td>
</tr>
<tr>
<td>UE</td>
<td>User Equipment</td>
</tr>
<tr>
<td>UE-RS</td>
<td>User Equipment-specific Reference Signal</td>
</tr>
<tr>
<td>UL</td>
<td>Uplink</td>
</tr>
</tbody>
</table>
## Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_c$</td>
<td>Average CBS area</td>
</tr>
<tr>
<td>$\hat{a}$</td>
<td>Grey model develop parameter</td>
</tr>
<tr>
<td>$\hat{a}_i$</td>
<td>$i^{th}$ polynomial coefficient of the approximated correlation function</td>
</tr>
<tr>
<td>$\hat{b}$</td>
<td>Grey model input</td>
</tr>
<tr>
<td>$\hat{b}$</td>
<td>Base of the approximated correlation function</td>
</tr>
<tr>
<td>$\hat{c}$</td>
<td>Polynomial order of the approximated correlation function</td>
</tr>
<tr>
<td>$C_{i,j}$</td>
<td>NL-based aggregated HO count from DBS$<em>{i}$ to DBS$</em>{j}$</td>
</tr>
<tr>
<td>$D$</td>
<td>Session duration</td>
</tr>
<tr>
<td>$D_r$</td>
<td>Residual session duration</td>
</tr>
<tr>
<td>$\mathcal{D}_s$</td>
<td>Doppler shift</td>
</tr>
<tr>
<td>$d_{s,\text{thr}}$</td>
<td>Prediction triggering threshold based on serving DBS distance</td>
</tr>
<tr>
<td>$d_{s,\text{thr},\text{UE}}$</td>
<td>UE-specific distance-based prediction triggering threshold</td>
</tr>
<tr>
<td>$G$</td>
<td>CN signalling reduction gain</td>
</tr>
<tr>
<td>$H(m)$</td>
<td>CFR at the $m^{th}$ subcarrier</td>
</tr>
<tr>
<td>$\tilde{H}(m)$</td>
<td>LS estimate of CFR at the $m^{th}$ subcarrier</td>
</tr>
<tr>
<td>$H(m,n)$</td>
<td>CFR at the $m^{th}$ subcarrier and the $n^{th}$ OFDM symbol</td>
</tr>
<tr>
<td>$\mathcal{H}$</td>
<td>HO signalling latency</td>
</tr>
<tr>
<td>$h(\cdot)$</td>
<td>Entropy of a probability vector</td>
</tr>
<tr>
<td>$h_{\text{thr}}$</td>
<td>Entropy threshold</td>
</tr>
<tr>
<td>$\mathbb{I}$</td>
<td>DTMC states’ indices</td>
</tr>
<tr>
<td>$I_p$</td>
<td>Predicted remaining time for HO in measurement gaps</td>
</tr>
<tr>
<td>$L$</td>
<td>Data plane backhaul latency</td>
</tr>
<tr>
<td>$L_p$</td>
<td>Propagation delay of the data plane backhaul path</td>
</tr>
<tr>
<td>$L_s$</td>
<td>Processing delay of the data plane backhaul path</td>
</tr>
<tr>
<td>$L_t$</td>
<td>Transmission delay of the data plane backhaul path</td>
</tr>
</tbody>
</table>
Symbols

$L_{\text{thr}}$ Maximum tolerable data plane backhaul latency

$M$ Total number of subcarriers

$
\mathcal{M}
$ DTMC state space

$N(m)$ Noise component at the $m^{th}$ subcarrier

$N_{c1}$ Average number of users in a CBS

$N_{c2}$ Average number of users leaving a CBS per unit of time

$N_f$ Theoretical maximum allowed FD pilot spacing for exponential PDP

$\tilde{N}_f$ Estimated maximum allowed FD pilot spacing

$\tilde{N}_t$ Estimated maximum allowed TD pilot spacing

$N_i$ List of first tier neighbours to DBS$_i$

$\hat{n}$ DTMC prediction set size

$\ddot{n}$ SS/SQ prediction window size per DBS

$\dddot{n}$ Maximum number of neighbouring DBSs monitored by the UE

$P_1^o$ Probability of not generating CN signalling in the CN-transparent HO scheme

$P_1^o$ Probability of not generating CN signalling in the CN-transparent HO scheme given that a CN signalling has already been generated

$P_2^o$ Probability of not generating CN signalling in the CN-visible HO scheme

$P_2^g$ Probability of not generating CN signalling in the CN-visible HO scheme given that a CN signalling has already been generated

$\mathbf{p}_0$ Initial distribution vector

$\mathbf{p}_k$ $k^{th}$ HO probability vector

$p_i$ Probability of being at DBS$_i$ after $k$ HOs

$\mathcal{P}_i$ Players set in stage $i$ of the game

$\mathcal{P}_i^2$ Total average power of the normalised channel impulse response

$Q_1$ Ratio between $P_1^o$ and $P_1^o$

$Q_2$ Ratio between $P_2^o$ and $P_2^o$

$q_n$ Neighbouring DBS transmit power

$q_s$ Serving DBS transmit power

$R_H(\cdot)$ FD CC or TD CC, with perfect channel knowledge

$\hat{R}_H(\cdot)$ FD CC or TD CC, with LS channel estimate

$\hat{R}_H(\cdot)$ FD CC or TD CC normalised with zero lag CC

$r_H(\cdot)$ Theoretical FD CC or TD CC

$r_H(\cdot,\cdot)$ Theoretical joint FD-TD CC
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_c$</td>
<td>CBS residence time</td>
</tr>
<tr>
<td>$R_{c,r}$</td>
<td>Residual CBS residence time</td>
</tr>
<tr>
<td>$R_d$</td>
<td>DBS residence time</td>
</tr>
<tr>
<td>$R_{d,r}$</td>
<td>Residual DBS residence time</td>
</tr>
<tr>
<td>$S_1$</td>
<td>Single inter-CBS HO CN signalling load in CDSA</td>
</tr>
<tr>
<td>$S_2$</td>
<td>Single HO CN signalling load in conventional architecture</td>
</tr>
<tr>
<td>$S_1$</td>
<td>Total CN signalling load in CN-transparent HO signalling scheme</td>
</tr>
<tr>
<td>$S_2$</td>
<td>Total CN signalling load in CN-visible HO signalling scheme</td>
</tr>
<tr>
<td>$s_{1,i}$</td>
<td>$i^{th}$ CN signalling state in CN-transparent HO signalling scheme</td>
</tr>
<tr>
<td>$T$</td>
<td>Transition probability matrix</td>
</tr>
<tr>
<td>$T$</td>
<td>OFDM symbol duration</td>
</tr>
<tr>
<td>$t_{i,j}$</td>
<td>Probability of direct transition from DBS$_i$ to DBS$_j$</td>
</tr>
<tr>
<td>$t_{min}$</td>
<td>Minimum HO transition probability to confirm the prediction from a history perspective</td>
</tr>
<tr>
<td>$U$</td>
<td>User density</td>
</tr>
<tr>
<td>$V$</td>
<td>User speed</td>
</tr>
<tr>
<td>$X(m)$</td>
<td>Transmitted signal at the $m^{th}$ subcarrier</td>
</tr>
<tr>
<td>$Y(m)$</td>
<td>Received signal at the $m^{th}$ subcarrier</td>
</tr>
<tr>
<td>$y^{(0)}(i)$</td>
<td>$i^{th}$ SS/SQ measurement before Grey AGO</td>
</tr>
<tr>
<td>$y^{(1)}(i)$</td>
<td>$i^{th}$ SS/SQ measurement after Grey AGO</td>
</tr>
<tr>
<td>$y_n^{(0)}$</td>
<td>SS/SQ of a neighbouring DBS</td>
</tr>
<tr>
<td>$y_p^{(0)}(i + j)$</td>
<td>$j^{th}$ predicted SS/SQ measurement referenced to the $i^{th}$ measured SS/SQ</td>
</tr>
<tr>
<td>$y_s^{(0)}$</td>
<td>SS/SQ of the serving DBS</td>
</tr>
<tr>
<td>$y_{s,thr}^{(0)}$</td>
<td>Prediction triggering threshold based on serving DBS signal</td>
</tr>
<tr>
<td>$Z(m)$</td>
<td>Channel estimation error at the $m^{th}$ subcarrier</td>
</tr>
<tr>
<td>$Z$</td>
<td>Data load</td>
</tr>
<tr>
<td>$z^{(1)}(i)$</td>
<td>Mean value of adjacent SS/SQ measurements after AGO</td>
</tr>
<tr>
<td>$\alpha_i$</td>
<td>Static data processing delay in node $i$</td>
</tr>
<tr>
<td>$\alpha_{i,j}$</td>
<td>Static data processing delay in repeater between node $i$ and node $j$</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>Data processing delay per bit in node $i$</td>
</tr>
<tr>
<td>$\beta_{i,j}$</td>
<td>Data processing delay per bit in repeater between node $i$ and node $j$</td>
</tr>
<tr>
<td>$b_i$</td>
<td>Binary variable of the initial transition from DBS$_i$</td>
</tr>
<tr>
<td>$\chi$</td>
<td>Distance-independent path loss component</td>
</tr>
<tr>
<td>$\Delta_m$</td>
<td>FD correlation lag in subcarriers</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>$\Delta_n$</td>
<td>TD correlation lag in OFDM symbols</td>
</tr>
<tr>
<td>$\delta_f$</td>
<td>Subcarrier spacing</td>
</tr>
<tr>
<td>$\delta_g$</td>
<td>Measurement gap</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Recent trajectory dependency parameter</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Span gradient</td>
</tr>
<tr>
<td>$\mathcal{G}$</td>
<td>Sum of independent gamma variables</td>
</tr>
<tr>
<td>$\gamma_i$</td>
<td>Gain of the $i^{th}$ path</td>
</tr>
<tr>
<td>$\kappa_i$</td>
<td>Shape parameter of gamma distributed processing delay in node $i$</td>
</tr>
<tr>
<td>$\kappa_{i,j}$</td>
<td>Shape parameter of gamma distributed processing delay in all repeaters between node $i$ and node $j$</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>Set of all nodes in the DP backhaul path</td>
</tr>
<tr>
<td>$\lambda_a$</td>
<td>AN density</td>
</tr>
<tr>
<td>$\lambda_c$</td>
<td>CBS density</td>
</tr>
<tr>
<td>$\lambda_d$</td>
<td>DBS density</td>
</tr>
<tr>
<td>$\lambda_{d_u}$</td>
<td>Upper bound of the DBS density under DP-BL constraints</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>Constant of proportionality between $\kappa_i$ and number of DBSs connected to node $i$</td>
</tr>
<tr>
<td>$\mu_{i,j}$</td>
<td>Constant of proportionality between $\kappa_{i,j}$ and number of DBSs connected to node $j$</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>UE angular span</td>
</tr>
<tr>
<td>$\Phi_a$</td>
<td>PPP of the AN distribution</td>
</tr>
<tr>
<td>$\Phi_c$</td>
<td>PPP of the CBS distribution</td>
</tr>
<tr>
<td>$\Phi_d$</td>
<td>PPP of the DBS distribution</td>
</tr>
<tr>
<td>$\psi$</td>
<td>Inter-site distance</td>
</tr>
<tr>
<td>$\sigma_{i,j}$</td>
<td>Scale parameter of Rayleigh distributed propagation delay between node $i$ and node $j$</td>
</tr>
<tr>
<td>$\Theta$</td>
<td>HO hysteresis</td>
</tr>
<tr>
<td>$\theta_i$</td>
<td>Scale parameter of gamma distributed processing delay in node $i$</td>
</tr>
<tr>
<td>$\tau_{rms}$</td>
<td>rms delay spread</td>
</tr>
<tr>
<td>$\Upsilon_f$</td>
<td>FD pilot correlation target</td>
</tr>
<tr>
<td>$\Upsilon_t$</td>
<td>TD pilot correlation target</td>
</tr>
<tr>
<td>$\varphi_{i,j}$</td>
<td>Propagation speed of the backhaul link connecting node $i$ to node $j$</td>
</tr>
<tr>
<td>$\varpi_{i,j}$</td>
<td>Transmission range of the backhaul link connecting node $i$ to node $j$</td>
</tr>
<tr>
<td>$\varepsilon_j$</td>
<td>Signalling processing latency in node $j$</td>
</tr>
</tbody>
</table>
Symbols

\( \varrho_{i,j} \)  
Number of repeaters between node \( i \) and node \( j \)

\( \varsigma_c \)  
Average CBS perimeter

\( \varrho_{i,j} \)  
Capacity of the link connecting node \( i \) and node \( j \)

\( \xi \)  
Path loss exponent

\( \zeta_{i,j} \)  
One way signalling transmission latency from node \( i \) to node \( j \)

\( \mathbb{E}[\cdot] \)  
Expectation

\( \text{erf}\{\cdot\} \)  
error function

\( F_W \)  
Cumulative distribution function of a random variable \( W \)

\( f_W \)  
Probability density function of a random variable \( W \)

\( \Gamma(\cdot) \)  
Gamma function

\( J_0(\cdot) \)  
Zero order Bessel function of the first kind

\( \mathcal{L}\{\cdot\} \)  
Laplace transform

\( \mathcal{L}^{-1}\{\cdot\} \)  
Inverse Laplace transform

\( \max_j \)  
maximum of arguments \( i \) s.t. constraint \( j \)

\( \min_j \)  
minimum of arguments \( i \) s.t. constraint \( j \)

\( \text{Prob}\{\cdot\} \)  
Probability of an event

\( \mathbb{R}^2 \)  
Two-dimensional Euclidean plane

*  
Complex conjugate

\( e \)  
Base of the natural logarithm

\( j \)  
Imaginary unit

\( \forall \)  
for all

\( \# \)  
Cardinality of a set

\( |\cdot| \)  
Magnitude

ceil \( \cdot \)  
Ceil operator
Chapter 1

Introduction

1.1 Background

Nowadays, requirements and performance bounds of fifth generation (5G) cellular systems are becoming of increasing interest in academia and industry fora. According to recent forecasts and worldwide discussions, an incremental advancement of current cellular systems, such as the long term evolution (LTE), may not be sufficient to satisfy the ambitious targets being identified for the 2020 era [1–3]. The proliferation of smart devices and the high dependency on mobile communications in everyday life have resulted in an exponentially increasing traffic demand. Among the possible techniques to overcome the capacity crunch problem, network densification is seen as the most promising solution. It has been estimated that 50 million base stations (BSs) will be deployed as soon as 2020 [4]. Although these estimations are debatable, they give an indication of the situation in the near future.

Such massive deployments raise several challenges from planning and cost perspectives. In this context, advanced concepts such as self-organising networks (SONs) can play a key role in minimising the planning and the management overhead. In addition, utilising existing infrastructure such as lamp posts can significantly reduce the deployment cost. Nevertheless, these dense deployment scenarios coupled with heterogeneity of network access points and new 5G use cases call for the design of efficient, flexible, scalable, sustainable and versatile cellular systems. These requirements are driven by
1.2 Motivation and Objectives

The conventional system was dimensioned based on a worst-case scenario to ensure acceptable performance for all users including those in severe conditions. In addition, it was designed to ensure ubiquitous coverage with an always present wireless communication channel irrespective of the spatio-temporal demand of service. Thus the evolution of current cellular generations is characterised by performance improvement without considering (or with a minimum attention to) other aspects. The overhead, environmental and economical impacts of this approach can be justified, to some extent, for the current density levels. However, several reports and recent forecasts on future deployment scenarios, use cases and energy consumption (EC) levels call for adopting a more signalling and energy conscious design. The latter requires an adaptive system
that is available at all locations and all times but becomes functional only when needed. This suggests a new clean slate system architecture with a logical separation between network and data access points, i.e., between the CP and the DP.

The main idea of the control/data separation architecture (CDSA) originates from the fact that only a small amount of signalling is required to enable ubiquitous coverage and network connectivity [5]. On the other hand, data transmission and its related signalling are needed on demand when there are active user equipment (UE). This calls for a two layer RAN architecture with a logical separation between:

- Network access and data transmission functionalities.
- Idle mode and active mode.
- Cell-specific/broadcast-type and UE-specific/unicast-type signalling.

In the CDSA, a continuous and reliable coverage layer is provided by control base stations (CBSs) at low frequency bands, where the large footprint ensures robust connectivity and mobility. The DP is supported by flexible, adaptive, high capacity and energy efficient data base stations (DBSs) that provide data transmission along with the necessary signalling. All UE are anchored to the CBS, while active UE are associated with both the CBS and the DBS in a dual connection mode [6]. A detailed mapping for the functionalities and the signalling supported by each layer is provided in Tables 2.1 and 2.2 of Chapter 2. With this configuration, the DBS is invisible to both detached and idle UE, and its on-demand connection with the active UE is established and assisted by the CBS. Expressed differently, the idle UE are connected with the CBS only. Thus the DBS layer is not constrained by the worst-case scenario and the DBS carrier can be switched off as long as it is not needed. This approach allows providing high data rate services under the umbrella of a coverage layer, which could open a wide range of benefits from signalling, energy, mobility and interference management perspectives.

It is worth mentioning that the CDSA is a relatively new concept in cellular domain, hence little work has been done to assess its superiority over the conventional architecture. In fact, most of this work provides a qualitative discussion rather than
1.3 Overview of Contributions

The main contributions of this thesis are summarised as follows:

1. A comprehensive survey of existing literature that investigates applications of the CDSA has been conducted. Several areas where the CDSA can overcome limitations of the conventional RAN architecture have identified and critically discussed. In addition, the research and technical challenges imposed by the CDSA have been pointed out, and the candidate solutions to tackle some of these challenges have been classified. Furthermore, the implementation aspects and enabling technologies of the CDSA have been compiled, along with the ongoing discussion in standardisation forums and international research projects. Based on this survey, three aspects have been identified as the most promising CDSA
benefits in futuristic dense deployment scenarios. These include: signalling overhead, mobility management and energy efficiency (EE). A wealth of literature is available for the latter aspect, i.e., EE of the CDSA. However, it has been found that little attention has been paid to investigate the signalling overhead and the mobility management in the context of RANs with CP/DP separation. This thesis serves as an introductory guide for futuristic cellular RANs with CP/DP separation, and it provides insights towards rethinking the RAN architecture to enable a green operation with efficient signalling and mobility management mechanisms. This contribution has been published in [7].

2. Physical layer (PL) signalling and DBS frame allocations under the CDSA are investigated. A pilot symbol-aided channel estimation approach is considered. To minimise the pilot and the PL signalling overhead, a correlation-based adaptive pilot signalling pattern has been proposed. This scheme considers DBSs with a multi-carrier air interface, and takes into account both frequency domain (FD) and time domain (TD) variations. It depends on estimating the actual channel correlation function under realistic and noisy channel conditions, rather than adopting the conventional worst-case design approach. In addition, the proposed adaptive scheme allows controlling the overhead/performance trade-off by including an adjustable correlation target. To assess potential gains of the proposed mechanism, the theoretical overhead of the adaptive signalling scheme is compared with the conventional worst-case pattern under several channel conditions and user speeds, and more than 90% reduction in the pilot overhead against the benchmark is achieved in local area (LA) and low mobility scenarios. In addition, the proposed scheme has been implemented in a link level simulator based on the LTE resource grid. The developed link level simulator provides a detailed modelling of the PL procedures such as channel encoding and decoding, modulation and demodulation, channel estimation and equalisation, resource grid generation, etc. Simulation results show that the adaptive scheme reduces the pilot overhead by 74%−78% w.r.t. the LTE pilot pattern without (or with a marginal) performance penalty. The proposed techniques in this contribution have been published in [8], [9] and [10].
3. Predictive and signalling efficient mobility management under CDSA with dual connectivity is investigated. To minimise the DBS HO latency and the associated signalling overhead, an advance HO-related DBS signalling for resource preparation has been proposed. These advance signalling mechanisms are enabled by HO prediction techniques since they are triggered before the actual HO criteria is satisfied. As a result, two DBS-level predictive HO schemes have been proposed. These include: a history-based predictive DBS HO signalling and a measurement-based context-aided predictive DBS HO signalling. The former predicts future HO events based on a Markov Chain modelling of previous HOs. It does not require maintaining memory consuming HO frequency tables, but rather it depends on an online learning processes. This prediction mechanism can also be used as an initial step in solving the DP holes problem for moving users in energy efficient CDSA implementation with DBS sleep modes. On the other hand, the measurement-based context-aided predictive HO scheme predicts both future DBS HO events along with the expected HO time. It combines radio frequency (RF) performance to physical proximity along with the UE context in terms of speed, direction and HO history. To minimise the processing and the storage requirements whilst improving the prediction performance, a user-specific prediction triggering threshold has been proposed. In both predictive schemes, a switching criteria between advance and conventional DBS signalling has been defined based on prediction entropy and successful HO probability, resulting in predictive HO schemes with two operation modes. The predictive techniques have been implemented in a system level simulator. The latter abstracts the PL procedures to reduce the simulator complexity whilst focusing on higher layer aspects related to mobility management, scheduling, multi-user handling, multi-cell deployment and interference management. The results show that the proposed schemes reduce the DBS HO signalling latency by a factor of $\frac{1}{3}$ w.r.t. the benchmark. The proposed techniques in this contribution have been published in [11], [12] and [13].

4. Signalling efficient mobility management has been further investigated from the CN perspective. Considering the dual connectivity feature of the CDSA, the
CBS has been exploited as a mobility anchor point for the user and as a DP anchor point for the DBSs. Based on this configuration, a CN-transparent HO scheme with minimal CN signalling overhead has been proposed. This scheme has been modelled analytically based on stochastic geometry. The proposed modelling approach allows analysing the impact of network density, user mobility and session characteristics on the HO-related CN signalling load. In addition, the DP backhaul latency (DP-BL) has been identified as the main CDSA challenge, particularly when the DP anchor point is moved to the CBS. Thus, the DP-BL has been modelled analytically and an upper bound for the network density has been derived to ensure that the DP-BL constraints are not violated. Furthermore, the CN-transparent HO signalling scheme has been integrated with the predictive HO signalling techniques to further minimise the HO latency. Numerical and simulation results show that the CDSA with CN-transparent HO reduces the HO-related CN signalling load by 70%–90% as compared with the conventional CN-visible HO approaches. In terms of HO latency, system level simulation results show that the integrated predictive and CN-transparent HO signalling scheme reduces the overall DBS HO latency by 60% w.r.t. the conventional HO approaches. The proposed techniques and modelling approach in this contribution have been published in [11], [14] and [15].

1.4 Thesis Structure and Outline

The remainder of this thesis is structured as follows:

First, Chapter 2 presents a holistic survey of existing literature on the CDSA for cellular RANs. As a starting point, it discusses the fundamentals, concepts, and general structure of the CDSA. Then, it points out limitations of the conventional architecture in futuristic deployment scenarios. In addition, it presents and critically discusses the work that has been done to investigate potential benefits of the CDSA, as well as its technical challenges and enabling technologies. Furthermore, an overview of standardisation proposals related to this research vision is provided.

Narrowing down, Chapter 3 discusses the applicability of adopting adaptive and dy-
namic frame allocation techniques in the CDSA, as opposed to the worst-case design of the conventional RAN architecture. It proposes an adaptive pilot signalling pattern for the DBS downlink (DL) frame. The proposed scheme considers multi-carrier systems with realistic and noisy channel conditions. By estimating the actual channel correlation function, the adaptive scheme is able to redistribute the pilots in both the FD and the TD to minimise the PL signalling overhead. The proposed scheme is compared against the conventional worst-case pattern in terms of performance and overhead for several user speed and channel delay spread values. In addition, a link level simulator is developed to evaluate the adaptive scheme against the LTE pilot design, and the simulation results are analysed and discussed.

Chapter 4 focuses on mobility and radio resource control (RRC) signalling in the CDSA, and proposes predictive mobility management at DBS-level. With the main objective of minimising the DBS HO latency and the associated air interface and RAN signalling, two predictive DBS HO signalling schemes are proposed. The system model of each scheme is described in Chapter 4 along with the HO procedure with and without mobility prediction. In addition, the prediction process and the HO signalling latency are formulated, and user-specific prediction parameters are derived. The proposed schemes are evaluated against the conventional HO procedures under several user mobility profiles and network parameters. Both numerical and system level simulation results are provided and discussed, and the impact of several parameters on the prediction and the signalling performance is analysed.

In Chapter 5, the CN signalling in the CDSA is discussed and a signalling efficient CN-transparent HO scheme is proposed. The system model of the proposed scheme is described and formulated. In addition, an analytical model based on stochastic geometry is developed to analyse the impact of network density, user mobility and session characteristics on the HO-related CN signalling load. Furthermore, the main challenge of the proposed scheme, i.e., the DP-BL, is modelled analytically and a latency constrained DBS deployment is proposed. The CN-transparent HO signalling scheme is evaluated against the conventional CN-visible HO signalling model under several user and network parameters. Numerical and simulation results are provided and discussed. Moreover, the CN-transparent HO signalling scheme is integrated with the predictive
HO signalling models, and simulation results of the integrated scheme are presented and discussed.

Finally, Chapter 6 provides the conclusions and summarises the main findings of this thesis. In addition, it underlines future research directions to extend the proposed techniques, as well as potential directions related to the CDSA research vision.

1.5 Publications

The research carried out during the course of this PhD has resulted in the following publications:

**Journal Articles**


**Book Chapter**

Conference Papers


1.6 Contributions Outside the Thesis

In addition to the work presented in this thesis, the authors are engaged in research and development activities for 5G system level simulators and integrated solution, as well as research activities on usage of cognitive radio in satellite/terrestrial spectrum sharing scenarios. In this regard, the following publications have been produced during the course of this PhD, but they have not been included in this thesis given that they are not entirely in line with its content.
1.6. Contributions Outside the Thesis


Chapter 2

Background and State-of-the-Art

This chapter provides a survey of existing literature that investigates the CDSA in cellular RAN domain. It identifies several areas where the CDSA can overcome limitations of the conventional architecture. In addition, the technical challenges imposed by the CDSA are identified along with candidate solutions and enabling technologies. Furthermore, some of the related ideas already under discussion in standardisation forums are presented. This chapter has been published in [7].

2.1 CDSA Concept and General Structure

2.1.1 Basic Operation

The key concept behind the CDSA is to separate the signals required for full coverage from those needed to support high data rate transmission. A few macro cells (MCs), also known as CBSs, provide network connectivity and support efficient RRC procedures. Within the CBS footprint, data services are provided by dedicated small cells (SCs) known as DBSs. As shown conceptually in Fig. 2.1, all UE are anchored to the CBS for basic connectivity services, e.g., system information, paging, channel requests, etc. while active UE are associated with both the CBS and the DBS in a dual connection mode. When the UE becomes active, e.g., starting a data session or receiving a call, the CBS selects the best serving DBS and establishes a high rate DBS-UE con-
2.1 CDSA Concept and General Structure

Connection through backhaul links. This approach comes with a range of benefits which are discussed in Sections 2.2–2.4.

Specifying functionalities of each plane is not trivial due to the fact that several functionalities may be needed to support a certain UE activity. For instance, cell reselection requires synchronisation and broadcast functionalities. In addition, a certain signal may be required by more than one network functionality such as the pilot signal, which is needed for synchronisation, paging, etc. [16]. In cellular domain, few separation schemes have been proposed to separate the CP from the DP. Based on network functionalities and a functionality-signal mapping, [16] proposed a separation scheme for the LTE by separating the functionalities required to support connectivity from those needed for data transmission. According to this scheme, the CBS supports synchronisation, broadcast, multicast, paging and RRC functionalities. On the other hand, the DBS supports unicast data transmission and synchronisation functionalities only. A similar approach has been followed in [17] for the global system for mobile communication (GSM). The authors of [18] argue that the control and the data channels are logically separated in current standards but they are mixed at the final stage to be transmitted by the same physical node. Thus, [18] proposed a separation scheme for LTE-Railway by mapping all logical control channels to a single physical channel that is transmitted by the CBS, while all logical traffic channels are mapped to a single physical traffic channel handled by the DBS. Table 2.1 maps network functionalities, while Table 2.2 maps UE states and shows frame allocations of the CBS and the DBS.

Figure 2.1: Control/Data separation architecture
Table 2.1: Functionality mapping in CDSA

<table>
<thead>
<tr>
<th>Functionality</th>
<th>CBS</th>
<th>DBS</th>
<th>Reason/Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell search</td>
<td>✓</td>
<td></td>
<td>Network access and connectivity are provided by the CBS only. DBSs can be switched off</td>
</tr>
<tr>
<td>System information</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paging</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multicast and broadcast data</td>
<td>✓</td>
<td>*</td>
<td>These services are supported by the CBS only to maximise the DBS transmission resources and sleep periods</td>
</tr>
<tr>
<td>Radio Resource Control</td>
<td>✓</td>
<td></td>
<td>Reduced handover overhead and improved mobility</td>
</tr>
<tr>
<td>Mobility management</td>
<td>✓</td>
<td></td>
<td>Performance by exploiting the wide CBS footprint</td>
</tr>
<tr>
<td>DBS selection</td>
<td>✓</td>
<td></td>
<td>Optimised network driven UE-DBS association based on a wide view of network status</td>
</tr>
<tr>
<td>Unicast data transmission</td>
<td>**</td>
<td>✓</td>
<td>One-to-one data transmission</td>
</tr>
<tr>
<td>Link adaptation</td>
<td>**</td>
<td>✓</td>
<td>These functionalities support data transmission and they require fast adaptation/response</td>
</tr>
<tr>
<td>Beam-forming</td>
<td>**</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

* When multicast/broadcast data is required by a clustered users of a specific DBS.
** When the CBS provides low rate services, e.g., voice or data transmission to high speed users.

Table 2.2: UE state and frame structure mapping in CDSA

<table>
<thead>
<tr>
<th>UE state</th>
<th>CBS</th>
<th>DBS</th>
<th>Reason/Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detached/Idle</td>
<td>✓</td>
<td></td>
<td>Idle UE maintain a single connection with the CBS only as long as they do not require data transmission. DBSs can be switched off</td>
</tr>
<tr>
<td>Active</td>
<td>✓</td>
<td>✓</td>
<td>Active UE maintain a dual connection: with the CBS for RRC and system information, and with the DBS for data transmission</td>
</tr>
<tr>
<td>Signal</td>
<td>CBS</td>
<td>DBS</td>
<td>Reason/Benefit</td>
</tr>
<tr>
<td>Synchronisation</td>
<td>✓</td>
<td>✓</td>
<td>Active UE need to synchronise with both carriers</td>
</tr>
<tr>
<td>Pilot</td>
<td>✓</td>
<td>✓</td>
<td>CBS and DBS could have different characteristics. Thus both frames need to contain pilot signal for channel estimation</td>
</tr>
<tr>
<td>Frame control</td>
<td>✓</td>
<td>✓</td>
<td>To specify the allocations within each frame</td>
</tr>
<tr>
<td>Paging</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Broadcast bearer</td>
<td>✓</td>
<td>*</td>
<td>The DBS frame does not need allocations for paging, broadcast and multicast bearer signals since they are provided by the CBS</td>
</tr>
<tr>
<td>Multicast bearer</td>
<td>✓</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Unicast bearer</td>
<td>**</td>
<td>✓</td>
<td>Most of the DBS resources are allocated to the unicast bearer</td>
</tr>
</tbody>
</table>

* When multicast/broadcast data is required by a clustered users of a specific DBS.
** When the CBS provides low rate services, e.g., voice or data transmission to high speed users

Although the CBS and the DBS support different functionalities, they need to coordinate and communicate with each other. The serving node selection decision requires the DBSs to exchange information about their current status, such as EC, congestion, etc., with the CBS. This coordination is also required to optimise scheduling, resource
allocation, interference management and mobility management. As an illustration, the CBS and the DBS may negotiate whether the UE will be served by the DBS (e.g., a low speed terminal) or whether it will be served at a low rate by the CBS only to minimise the mobility overhead in high speed scenarios. Tight collaboration and excessive signalling exchange between the CBS and the DBS provide reliable, robust and updated information. However this increases the system overhead [17] and requires an ideal backhaul connection, i.e., high throughput and low latency [19]. On the other hand, a low rate CBS/DBS signalling relaxes the backhaul requirements and reduces the overhead but it may result into unreliable or out-of-date information. This calls for the design of efficient CBS/DBS signalling mechanisms whilst balancing the backhaul signalling periodicity/overhead trade-off. The concept of over-the-air (OTA) signalling [20], [21] has been recently proposed as an alternative solution to avoid using the conventional direct backhaul networks. When the CBS takes full control of scheduling functionalities, the OTA signalling concept allows the DBSs to overhear the grants issued by the CBS [22]. This alleviates the need for an ideal backhaul, however, a robust signalling design is required for interference avoidance.

It is worth mentioning that the CDSA is a new concept in cellular domain although a comparable approach has been proposed earlier for other systems, such as sensor networks [23], [24]. Thus, its operation and implementation aspects are currently being studied in several international research projects. These include:

- **Beyond Cellular Green Generation (BCG²)** [25]

  This is a project of the GreenTouch Consortium with a primary target of improving EE of cellular systems. It focuses on benefits of the CDSA from an energy perspective and proposes a cell on-demand approach. In the latter, the DBSs are switched on and off according to traffic variations without affecting the basic connectivity service provided by the CBS. However, such an operation raises several challenges as discussed in Section 2.5. In particular, BCG² tackles the problems of context information detection, serving node selection and management of interaction between the CP and the DP.

- **Toward Green 5G Mobile Networks (5grEEn)** [26]
As with BCG\(^2\), 5grEEn focuses on designing green 5G cellular networks with a logical separation between idle mode functions and data transmission services (i.e., CP/DP separation). It investigates the usage of massive reconfigurable antennas with dynamic cell structuring to optimally reshape the DP coverage. Such techniques adapt the network to traffic variations and allow increasing the inter-site distance, thus reducing EC and improving overall efficiency of the network. In addition, 5grEEn investigates the impact of the CDSA on network deployment strategies and possible backhauling solutions.

- **Millimetre-Wave Evolution for Backhaul and Access (MiWEBA) [27]**
  This is a joint European Japanese research project with a primary target of extending cellular systems capacity by exploiting the millimetre wave (mm-wave) band, i.e., above 60 GHz. MiWEBA integrates mm-wave SCs into conventional cellular systems, and utilises the CDSA to overcome coverage restrictions of the mm-wave link. The network architecture consists of MCs placed on rooftops to provide the basic connectivity service at conventional cellular bands. Data services are provided by mm-wave SCs that are deployed within the MC footprint. Depending on the deployment scenario, the MiWEBA project investigates whether the CP and the DP should be logically and physically separated (i.e., provided by separate physical nodes) or whether it is more feasible to adopt a logical separation only (i.e., control and data interfaces are hosted in the same node). Several key performance indicators (KPIs) are considered in analysing this trade-off such as data channel acquisition delay, data session retainability, EC and signalling overhead.

- **Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS) [28]**
  The FP7 research project METIS defines, investigates, characterises and models a potential 5G air interface and considers the CDSA as a candidate interference management technique. Targeting a minimal inter-cell interference (ICI), METIS exploits the wider view of the CBS that controls power and resource allocations of the DBSs under its footprint, by using centralised interference-aware
2.1. CDSA Concept and General Structure

2.1.2 CDSA and Software-Defined Networking

Software-defined networking (SDN) is an emerging concept that decouples the CP and the DP by separating control decision entities from control action enforcement elements. Although the basic idea of the SDN sounds similar to the cellular CDSA, these two concepts should not be confused with each other. In SDN, CP means the decision scheduling mechanisms [19]. Contextual information and signal-to-noise ratio (SNR) databases are used for channel quality prediction. In addition, METIS investigates the usage of carrier aggregation to enable a seamless implementation of the CDSA in current standards [19]. Fig. 2.2 shows a high level diagram of the CDSA implementation aspects and potential benefits that are investigated in these projects.
2.1. CDSA Concept and General Structure

Table 2.3: Comparison between SDN and CDSA

<table>
<thead>
<tr>
<th>Comparison criteria</th>
<th>SDN</th>
<th>CDSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope</td>
<td>CN</td>
<td>RAN</td>
</tr>
<tr>
<td>Network elements</td>
<td>routers</td>
<td>DBSs</td>
</tr>
<tr>
<td>Central units</td>
<td>Central traffic controllers</td>
<td>CBS</td>
</tr>
<tr>
<td>Unique advantages</td>
<td>Software upgrade, technology agonistic, softly defined capacity</td>
<td>Mobility robustness, easy interference management</td>
</tr>
<tr>
<td>Unique challenges</td>
<td>Delay</td>
<td>Functionality/Signalling separation, backhaul networks, frame structure</td>
</tr>
<tr>
<td>Common advantages</td>
<td>Energy saving, cost saving, efficient resource management</td>
<td>Single point of failure</td>
</tr>
<tr>
<td>Common challenges</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

makers that determine where and how the traffic should be sent, while DP refers to the system that forwards the packets according to the decision taken by the CP. SDN allows cellular networks to be flexible and reconfigurable and it simplifies network management procedures [29]. This is realised by moving the CP to a software application often called controller, resulting into a programmable network. In cellular CDSA, the CP includes decision making entities in addition to (most of) the network-UE signalling that is related to the service being requested/provided by/to the UE. This signalling includes RRC connection establishment and maintenance commands, scheduling information, etc. without which data transmission and seamless connectivity cannot be supported.

As discussed in Section 2.4, the CDSA allows a paradigm shift towards almost centralised control functionalities. Aligned with this trend, the SDN suggests a centralised controller to enable control decisions to be taken based on a wider view of network status and parameters [29]. In other words, the SDN and the CDSA have the same physical realisation: moving towards a centralised CP [30]. SDN and CDSA are closely related concepts in the sense. In both architectures, intelligence is partially or completely removed from most of the nodes in the network to be concentrated in fewer central nodes. This results in cost saving, higher performance and resource efficiency. SDN is manifestation of the above idea in CN, whereas CDSA implements the same idea in RAN. A comparison between the SDN and the CDSA concepts is provided in Table 2.3. Recent studies have proposed integrating the SDN and the CDSA, [31] refers to such integration as Soft-RAN. In the latter, the SDN concept is adopted to abstract
all BSs as a virtual big BS (analogous to the CBS) that hosts a centralised CP for radio elements (analogues to the DBSs). Following a similar approach, [32] proposes a programmable 5G CP where connectivity is provided as a service application running in the controller. In addition, the SDN/CDSA integration has been investigated in [33], where a two layer 5G network architecture has been proposed.

A similar architecture has been proposed in the FP7 CROWD\(^1\) project by combining the CDSA with the SDN. It follows the classical SDN approach of using a centralised controller whilst reducing signalling overhead by terminating some of the control information in local controllers, resulting in a hybrid centralised/distributed control functionalities [34]. The CP is implemented in a software application handled by the local controllers that are hosted in RAN elements and they are used for fast and fine grained control functionalities. Several local controllers are connected to a regional controller hosted in a data centre, which is used for slower, long time scale control operations [35]. The CDSA has been adopted by directing control path of the LTE to the local controllers, while the data path goes to a distributed mobility management entity (MME) gateway. The reader is referred to [34] and [35] for detailed description of the CROWD architecture.

The authors of [30] and [36] integrate the SDN concept with the BCG\(^2\) architecture, and they argue that the CDSA requires redesigning current network hardware components. Thus the SDN is seen as an enabling technology that could allow a feasible and cost-efficient CDSA implementation. In addition, the SDN offers a technology-agnostic CP by allowing the control decisions and commands to be taken at a technology-agnostic level of abstraction [34]. This feature is of great importance when the CBS manages DPs of several operators in infrastructure and/or spectrum sharing scenarios [37]. Furthermore, the SDN enables the CDSA applications related to network-driven resource selection. This can be done by implementing an application that collects information on network status and UE context, and then executes optimisation functions to dynamically associate the UE with the best serving DBS. The optimisation function can have an objective of increasing the EE (e.g., associate the UE with a small subset of the DBSs and switch off other DBSs), balancing the network load (e.g., offload some

---

\(^1\)Connectivity management for eneRgy Optimised Wireless Dense networks.

\(^2\)Bottom-up control and global optimisation
2.2 System Capacity

The on-going trend towards Internet of Things (IoT) and Machine-to-Machine (M2M) communications will increase the number of connected devices in 2020 by at least a factor of 10 compared with 2009 figures [38]. With an average user data rate increase by 50–100 fold to reach a peak target of 10 Gbps [4], it is estimated that 1000-fold increase in the capacity will be required in the next decade [39].

2.2.1 Capacity Expansion Mechanisms

To satisfy this demand, several techniques are being studied and standardised such as massive and enhanced multiple input multiple output (MIMO) mechanisms, beamforming, carrier aggregation, coordinated multipoint (CoMP), bandwidth (BW) expansion and SC deployments. Based on [4] estimates, the capacity cube of Fig. 2.3 groups these techniques into three main categories: spectrum extension, spectrum efficiency and network density. Enhanced MIMO and beam-forming techniques improve the spectral efficiency (SE) but they may not be sufficient to achieve the ambitious 1000-fold
capacity target. MIMO increases the over-the-air throughout by a factor of the number of antennas. However, it introduces an additional pilot overhead which decreases the actual gains in the goodput (i.e., the application level throughput). CoMP techniques depend on BSs cooperation to enhance the PL performance and to mitigate the ICI for cell edge users. Although these mechanisms may be necessary in dense deployment scenarios, they cannot achieve the 1000-fold capacity increase [40].

On the other hand, the proportional relationship between the BW and the capacity depicted by Shannon’s formula [41] indicates that wider BWs give higher capacity. In this direction, carrier aggregation has been proposed in [42] for LTE-Advanced systems, where two or more carriers are aggregated together resulting into a maximum aggregated BW of 100 MHz [43]. However, this technique is limited by the allocated BW and hence wider BW allocations are required for future systems. Two solutions to this problem are identified:

- Spectrum sharing.
- New spectrum exploitation.

The former shares the same portion of the spectrum between different operators under specific regulation and coordination rules, while the latter suggests exploiting new frequency bands. Nonetheless, scarcity of spectrum resources in low frequency bands requires exploiting higher bands where free portions of the spectrum are available. As a result, regulatory and standardisation bodies are considering high frequency bands, such as 3.4 – 3.6 GHz and above 6 GHz, as the main candidates for future cellular systems [44]. In addition, the mm-wave bands (i.e., above 60 GHz) are being considered as a spectrum extension solution to satisfy the increasing capacity demand. Nonetheless, the high propagation loss of such bands limits their usability to LA and short range communications only.

Network densification allows spatial reuse of spectrum resources by reducing the cell size. The idea originates from the fact that deploying several SCs instead of one MC i.e., cell splitting, allows resource reuse across the cells. For example, in frequency reuse of one systems splitting a MC into two SCs could result into doubling the capacity. Hence
it can be said that spectrum extension and network densification are highly correlated. In particular, a dense deployment of SCs has been accepted to be the most promising solution to satisfy capacity demands of future cellular systems [45], [46]. As a result, more SCs are being deployed within the MC coverage to offload some of the users associated with the latter. This is referred to as heterogeneous networks (HetNets), which is being considered for LTE-Advanced and beyond [47].

### 2.2.2 Heterogeneous Networks

In conventional HetNets, the MCs and the SCs are deployed in the same frequency band [48], thus inter-layer interference mitigation techniques such as spectrum splitting or almost blank subframes are required. However, these techniques may degrade the achievable capacity because they segment the resources between the layers either in FD or TD. This suggests a frequency-separated deployment, where each layer is deployed in a separate frequency band to avoid the inter-layer interference. Nonetheless, it is worth emphasising that the frequency-separated deployment also results into resource splitting loss.

The CDSA is aligned with the frequency-separated deployment approach. Since the CBS provides low-rate/long-range coverage services, it can use the existing low frequency bands that offer good propagation capabilities. On the other hand, the capacity hungry plane, i.e., the DBSs, can operate at high frequency bands that offer more spectrum resources and higher capacity. This approach is being investigated by several operators and research projects as a novel solution for future cellular systems. In this direction, the Phantom Cell Concept (PCC) was proposed in [6], where coverage and data services are provided at low and high frequency bands respectively. The models developed in [49] show that SE of the PCC outperforms the SE offered by conventional HetNets.

In [50], a similar architecture called Cloud-HetNet has been proposed with a primary target of extending the capacity. In Cloud-HetNet, all cells (i.e., MCs and SCs) are connected to a Cloud-RAN in a star topology and they act as radio resource heads. The Cloud-RAN is the brain of this architecture, where network and medium access
2.2. System Capacity

control layer functions and part of the baseband processing are performed in a centralised manner. Based on the Cloud-RAN, the MC can handle the CP of all users for mobility and cell discovery, while the DP is supported by the SCs. An interesting finding of [50] is that operating the DP at the 3 GHz band (where 100 MHz BW is still available) provides more capacity than the 60 GHz band (where 2.16 GHz BW is available) when the traffic load is low and vice versa. This is because [50] defines the capacity as the minimum of the achievable throughput and the traffic rate. Although the 3 GHz band offers limited BW, it can satisfy low traffic rate demands. Thus the wider BW offered by the 60 GHz band does not provide additional capacity gains and the reduced coverage minimises the achieved capacity in low load situations. In addition, the absence of cell-specific reference signals (CRSs) in the DP and the flexibility in switching off the DBSs reduce the DP ICI, which in turn increases the signal-to-interference-plus-noise ratio (SINR). According to Shannon’s capacity formula, the latter can be translated into an increased capacity. Moreover, a higher SINR allows using high order modulation and coding schemes that provide high data rates.

2.2.3 Scalability and Reconfiguration

Heterogeneity of future networks will create high variations over spatial, time and frequency domains due to mobility, variable-rate applications and SC deployments. This requires flexible, cost efficient and reconfigurable networks that are able to adapt to such variations. In this regard, network adaptation and reconfiguration might be easily performed in the CDSA with relaxed constraints. For instance, the DP can be flexibly scaled without coverage restrictions. Thus network operators can start by deploying DBSs to satisfy the current demand only and gradually add capacity when and where it is needed. In [51], such a scalable architecture is referred to as fusion network where a host layer guarantees the connectivity while a scalable and flexible boosting layer provides on-demand high data rate services.

In dense deployment scenarios, traffic tendency of each cell will be prone to high fluctuations e.g., a cell may be characterised by an asymmetric uplink (UL)/DL traffic. In such cases, assigning static or semi-static resources for the UL and the DL could
result into resource wastage. This requires flexible (re)allocation schemes to ensure efficient usage of spectrum resources. One of these schemes is dynamic time division duplex (TDD) that shares all the time slots between the UL and the DL with flexible slot reconfiguration [52]. Semi-static variations of this technique have already been implemented in current standard, for example the LTE defines seven UL/DL slot configurations [53]. However the mandatory transmission of the CRS and other periodic signals limits these techniques in the conventional architecture.

On the other hand, the absence of periodic CRS/broadcast-type signalling in DBSs of the CDSA could offer a flexible implementation of dynamic TDD. Nonetheless, interference coordination between neighbouring DBSs may be required because using different UL/DL configurations in different cells implies that there could be UL-to-DL and DL-to-UL ICI in addition to the classical UL-to-UL and DL-to-DL ICI [54]. To solve this problem, a hybrid frequency division duplex (FDD) and TDD coordination scheme (hybrid FDD-TDD) has been proposed in [54] as an interference coordination technique for DBSs utilising dynamic TDD. This scheme avoids the UL-to-DL and the DL-to-UL ICI by scheduling the UL and the DL for each user in different carriers and at different subframes.

2.3 Energy Efficiency

Current cellular systems have been developed and evolved with a primary focus on performance improvement without considering energy aspects [55]. Nowadays, the information and communication technology sector contributes 3% to the global EC and generates 2% of the worldwide CO$_2$ emissions [56], with recent forecasts for doubling this contribution every five years [57]. Thus, neglecting the energy dimension in designing 5G cellular systems will cause them to encounter several environmental and economical problems. In wireless systems, most of the energy is consumed by radio interface components. Precisely, more than 80% of the access network power in cellular systems is consumed by the BSs [58]. As a result, minimising EC of the access network is the best way to conform to the general trend of sustainable and green communications, as well as to cut the energy bill.
2.3. Energy Efficiency

Conventional cellular systems consume high power even in low traffic situations due to the “always-on” service approach adopted in these systems. The results of the EARTH$^2$ project reported in [59] show that EC of the LTE is almost insensitive to traffic load and is dominated by unnecessary overhead transmission and idle mode signalling, see for example a typical power profile of pico BSs in Fig. 2.4. The most power consuming component in the BS i.e., the power amplifier (PA) [59], can also be considered as one of the contributors to this load-independent energy profile. It consumes 55%–60% of the overall power consumption of macro BSs, and 30%–36% of the power consumption of pico BSs [59], [60]. The most efficient operating point for the PA is near the saturation region. However, non-constant envelope modulation schemes require the PA to operate in the linear region to avoid non-linear distortion and adjacent channel interference. As a result, a back-off from the saturation is usually employed in commercial PAs which reduces the PA efficiency to 5%–40% [61].

Advanced techniques such as clipping, pre-distortion and Doherty PA are used to increase the linear region of the PA in macro BSs. Thus the power consumption of the latter scales, to some extent, with the traffic load. In pico BSs, however, such advanced techniques are not used and the PA is operated with a lower efficiency. This efficiency becomes even worse at medium and low load resulting in a load-independent energy

\footnote{Energy Aware Radio and neTwork technologies. http://www.ict-earth.eu}
consumption profile in pico BS PAs [62]. This suggests introducing scalability into the PA design by using advanced techniques, such as multi-stage PA [63], that support dynamic power management. Although these techniques improve the EC profile, such component level optimisation does not overcome the baseline power consumed by active PAs [64].

This EC profile can be justified in high traffic scenarios. However, today’s cellular networks operate in a low load regime. Currently, the average BS utilisation is less than 10% for 45% of the time [65] and [66] estimates that up to 97% of orthogonal frequency division multiplexing (OFDM) wireless resources are not used, with 50% of the traffic being carried by 15% of the deployed BSs [67]. Given the load-insensitive EC profile, it can be concluded that EE of current cellular systems is generally poor.

2.3.1 Conventional Energy Saving Techniques

Minimising the EC requires exploiting the spare capacity by adapting the network to the actual traffic load. In this context, several energy saving techniques, such as discontinuous transmission, MIMO muting and cell wilting and blossoming, have been proposed.

- Discontinuous transmission: provides energy saving in TD by switching off some of the BS components such as the PA during unoccupied subframes\(^3\). Nonetheless, the mandatory transmission of CRSs limits the sleep periods in this technique [68].

- MIMO muting: provides energy saving in spatial domain by reducing the number of active antennas [68]. This technique is of great importance since MIMO systems increase the EC significantly due to the large number of PAs and the complex processing for multiplexing and diversity gains. However, MIMO muting requires fast adaptation to satisfy coverage and performance requirements [69].

- Cell wilting and blossoming: this technique exploits the fact that energy loss is proportional to propagation distance [70] by adapting the cell size to traffic profile.

\(^3\)In the LTE, the radio frame consists of ten subframes. Each subframe is divided into two time slots of 0.5 ms each.
Cell wilting lessens the pilot power in off-peak periods to allow soft reduction of the coverage. When the traffic demand rises, cell blossoming increases the pilot power [71]. However, cell size adaptation may not be timely enough to respond to rapid variations in data traffic. In addition, it requires careful HO management because the reduced overlap area may increase the call drop rates [68].

### 2.3.2 Coverage Restrictions

Achieving a breakthrough in energy saving requires a paradigm shift towards on-demand systems by switching off a subset of the BSs during off-peak periods [72]. Although such wide network adaptation will result in a significant saving in energy, it may not be feasible with the conventional cellular architecture due to the tight coupling between coverage and data services. In other words, coverage constraints are considered as the main source of energy inefficiency. In this regard, several techniques have been proposed to preserve the coverage; power control techniques, such as cell zooming, can be employed to increase the power of some BSs when other BSs are switched off [73]. Despite its potential gains, this technique may not guarantee full coverage and provides poor performance to cell edge users due to the increased ICI between active BSs with an extended coverage [73]. Recently, suboptimal sleep mode mechanisms have been proposed for SCs in the third generation partnership project (3GPP). In these mechanisms, either the RF receiver chain of the BS has to be kept on to receive signalling to switch on the BS, or the RF transmitting chain has to be turned on periodically to transmit beacon signals [74], [75].

Multi-hop relay has been proposed in [76] to allow other terminals to relay the traffic of UE in the vicinity of a switched off BS. Although this technique does not increase the ICI, finding suitable relays is a challenging task [76], and the received signal at the UE can be very poor depending on location and capabilities of the relays. BSs cooperation techniques, such as CoMP, can also be used to provide coverage to UE when their nearest BS is switched off. Although the joint transmission of several BSs boosts the received signal at the UE, this technique guarantees neither performance nor full coverage for all affected users. Based on this discussion, it can be concluded that the
conventional cellular architecture where basic coverage functionalities and data transmission services are provided by the same physical node offers limited opportunities for energy saving. In addition, most of the standardised/proposed techniques are limited by the coverage constraints as well as the mandatory transmission of CRSs.

2.3.3 Energy Saving in CDSA

Separating the CP from the DP allows flexible adaptation opportunities without breaking the anywhere/anytime service paradigm. In the CDSA, the basic coverage is provided by a few CBSs, while data transmission is supported by DBSs as shown in Figs. 2.1 and 2.2. Hence, adapting the DBSs to traffic load does not affect the coverage provided by the CP. Expressed differently, the CDSA could allow a paradigm shift towards on-demand always-available systems that scale the EC with the traffic load whilst maintaining a full connectivity coverage.

Considering the EARTH 2020 traffic model, [25] shows that the flexible opportunities for DBS on/off operation achieve up to four times higher EE compared with legacy systems. Reference [16] incorporates the DBS sleep opportunities along with the reduction in control signalling, and shows that such an architecture can save up to one third of the energy in urban deployment scenarios whilst scaling the EC with the traffic load. The feasibility study reported in [77] indicates that potential energy gains of the CDSA will be much higher in low utilisation and dense deployment scenarios. However, this study does not consider the facts that each DBS has a finite capacity and the instantaneous utilisation of the DBS affects its ability to serve other users. In addition, decoupling the CP from the DP allows flexibility in reshaping the coverage of the DBS (i.e., cell re-structuring) without affecting the underlay CP coverage. In contrast to the conventional architecture, the DBS does not transmit CRSs [16]. Thus the DBS can be considered as a UE-specific resource that dynamically transmits the data in directions towards the active UE only. Considering this feature, [50] proposed a dynamic cell structuring mechanism by using large-scale CoMP. In this technique, a cluster of DBSs is dynamically created around hotspots by controlling beam directions of each DBS. Furthermore, cell wilting and blossoming can be easily realised in the
2.3. Energy Efficiency

![Graph showing Energy Efficiency and Spectral Efficiency vs DBS density]

Figure 2.5: Energy and spectral efficiencies vs DBS density, based on [81] models

DP with relaxed HO constraints when mobility management is delegated to the CBS. These flexible opportunities for power adjustment and beam-forming result into a high gain which can be translated into an increase in the link level EE [26].

It is well understood from a number of recent studies [78–80] that in conventional network operational point for EE and SE are not the same. Network operator has to choose between the two KPIs while designing a network. One method to optimise this trade-off dynamically, while taking into account spatio-temporal variation of traffic demand, is to switch on and off the BS. However, conventional cellular networks are not designed for frequent switching on and off. Whereas CDSA, as explained above has all features needed to perform dynamic on and off switching with high agility. A very recent study in [81] has investigated the technical benefits of the CDSA in terms of both SE and EE. An interesting finding from [81] is summarised in plot shown in Fig. 2.5. It can be seen from Fig. 2.5 that the deployment density of DBS that yields maximum EE and deployment density of DBS that yields maximum SE are not the same. This again reinforces the conclusions from EE vs SE studies on conventional networks [78–80]. However, unlike the conventional architecture, where dynamically changing density of BS is not administratively as well as technically feasible, due to intrinsic decoupled design, in the CDSA effective density of the DBSs can be orchestrated in self-organising fashion with much better administrative and technical ease. This dynamic adaptation of effective DBS density can then allow to choose desired operational point between
EE and SE by maintaining optimal effective DBS density, while taking into account spatio-temporal traffic demand.

In addition to the flexibility of trading EE and SE, a recent study in [82] has also shown that the CDSA can offer better SE mainly because of selection diversity that stems from large number of DBSs. Opportunities for centralised interference coordination, as discussed in Section 2.4.1, is another feature of the CDSA that can yield better SE as compared to conventional networks. The high SE means the transmission will be done quickly which increases the DBS quiet period i.e., more time for the DBS to operate in sleep mode [26]. In addition, better EE in the CDSA mainly comes from low power DBSs, ability to switch off DBSs, lower propagation losses due to smaller UE-DBS distance, and opportunity for centralised self-organising EE functions to switch off and on suboptimal used DBSs in conjunction with load balancing SONs [83]. Furthermore, the low-rate/long-range services provided by the CP allow using low order constant envelope modulation, such as BPSK\(^4\) and QPSK\(^5\), in the CBS. Thus the PA of the CBS can operate at saturation without non-linearity problems, which improves the PA efficiency and hence the EE.

2.4 Centralised Control Plane

This section surveys and discusses the benefits of using a CP with central scheduling and decision entity rather than fully distributed decisions at the DP. In particular, it focuses on applications related to interference, resource and mobility management.

2.4.1 Interference Management

Interference control is a major concern in cellular systems especially those adopting a frequency reuse of one. To cope with this issue, several interference mitigation techniques have been standardised in the LTE, such as resource partitioning between the cells and resource muting during CRS transmission of other antenna ports, as shown in Fig. 2.6. Other advanced interference management strategies have been considered,

\(^{4}\) Bipolar Phase Shift Keying.
\(^{5}\) Quadrature Phase Shift Keying.
Figure 2.6: LTE CRS pattern of four antenna ports. Acronym PDCCH: Physical downlink control channel. PDSCH: Physical downlink shared channel.

such as:

- Slowly-adaptive interference management [47].
- Enhanced ICI coordination (eICIC) [84].
- Autonomous component carrier selection [85].

eICIC mitigates the ICI for cell edge users by coordinating network resources in time, frequency and power domains [84]. To cope with the ICI in range expansion zones of HetNets, an extension of eICIC that complements time domain resource partitioning techniques, such as almost blank subframes, with non-linear interference cancelling receiver processing has been proposed in [86] and [87]. With dense deployment of SCs, these techniques offer limited flexibility and may not be responsive to rapid traffic variations [88]. In addition, the interference coordination will be problematic due to the increased number of interferers, and a centralised coordinator may be required to control the resource usage among different cells.

The CDSA offers flexibility in this context because the CP can play the role of the centralised coordinator. In [19], different approaches to control the ICI in control/data separation scenarios have been identified. In one scenario, the CBS fully controls the scheduling for the DBSs, which overhear the grants issued by the former. In this case, the UE request resources from the CBS, which associates each user with the best
serving DBS and schedules the users of neighbouring DBSs on different resources. This approach does not require a backhaul signalling between the CBS and the DBS, but it may introduce delay in the DL scheduling [19]. Another approach is to maintain the scheduling functionalities at the DP with scheduling constraints being defined by the CP. However, such an approach generates additional signalling between the CBS and the DBS.

### 2.4.2 Resource Management

Traditionally, cellular users camp on the network by selecting the BS that offers the strongest signal. Thus cell (re)selection is mainly UE driven with a limited control by the network. For instance, the network may use offset parameters, usually called bias or cell range extension factor, to privilege some cells [89]. Since the cell (re)selection does not require resource assignment, the UE driven approach can be justified in this case. However, the active UE are assigned resources by the same cell initially selected by the user, which puts constraints on the resource management and optimisation process, e.g., the resources have to be assigned by this cell only without a global view of the network.

In the CDSA, the initial access procedure can be based on the signal strength (SS), thus the cell (re)selection could be UE driven as in the conventional architecture. However, when the UE requests resources for data transmission, the CBS selects the best serving DBS (or a group of candidate DBSs) with a wide view of network status and parameters such as EC, congestion, performance requirements, etc. This allows a transition from almost distributed and UE driven to almost centralised and network driven radio resource management mechanisms and optimises the resource allocation process. A centralised resource management entity (RME) has been proposed in [90] for cellular networks with a CP/DP separation. The main responsibility of the RME is to select the best serving DBS based on network status and context information collected from the UE and the network nodes. The RME is accessible to both the CBS and the DBS and several trade-offs are identified in [90] to optimise the RME decision. A conceptual hierarchy of CDSA with RME is shown in Fig. 2.7.
2.4. Centralised Control Plane

The centralised resource management schemes do not only optimise the resource selection decision, but also they could help in balancing the network load. In dense SC scenarios, the number of active UE per cell is expected to be small. Thus each cell will be characterised by a highly fluctuating traffic profile [91]. In this case, the CBS (or a separate RME) can determine an average load threshold for each cell and then exposes only the appropriate set of DBSs to the UE in order to balance the traffic load. Another approach is to allow the UE to conduct measurements of surrounding DBSs based on which the CBS (or the RME) can allocate the serving DBS. From another perspective, the flexible opportunities for power adjustment in the DP can be realised to temporarily reshape the coverage of a low utilised DBS to overlap with a neighbouring congested DBS, thus offloading the latter and balancing the network load. In current systems, such coverage reshaping is very limited due to CRS interference as well as the constraints imposed by the planned coverage [6].

2.4.3 Mobility Management

As the cell size decreases, mobility management becomes complex because the HOs will happen frequently even for low mobility users. In the conventional cellular architecture, the HO procedure includes transferring all channels (i.e., control and data) from one BS to another with a significant signalling load [92]. With a frequent HO rate, signalling overhead and call drop rates will increase significantly, which could degrade the quality
of experience. On the contrary, the CDSA could offer robust HO procedures. In [93], the HO failure (HOF) and the radio link failure (RLF) rates are used as KPIs to analyse mobility performance of the CDSA. The RLF rate is defined as the average number of RLF occurrence per UE per second, where RLF is triggered when the DL SINR is below a certain threshold (typically $-8$ dB) and stays below $-6$ dB for at least 1 s [94]. On the other hand, the HOF rate is defined as the ratio between the number of HOF and the total number of HO attempts, where the HOF is triggered when the RLF occurs during the HO execution time [94]. The authors of [93] argue that the UE is always anchored to a MC (i.e., the CBS), thus the RLF and the HOF rates reflect the macro layer mobility. System level simulation results show that the RLF and the HOF rates of the CDSA are roughly 0.6%. The authors of [95] follow a similar approach and show that the HOF rate in the CDSA is 40% less than the HOF rate in the conventional architecture. It is worth mentioning that [93] and [95] assume that the MCs and the SCs are deployed in separate frequency bands. Thus the inter-layer interference is ignored. In co-channel deployments, however, this interference might be significant, which could degrade the SINR and increase the RLF and the HOF rates. This indicates that deploying the CP and the DP in separate frequency bands might be more appropriate from a mobility point of view.

Context information such as mobility history can play an important role in optimising the RRC and the 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 [5], [90]. In addition, predicting the user’s trajectory, i.e., a sequence of DBSs that the user will visit them, allows the CBS to make advance HO decisions. Thus each candidate DBS in the user’s path can prepare and reserve resources in advance, which in turn could relax the HO requirements and minimise the interruption time. Moreover, the CBS can provide the candidate DBSs with some information about the UE, such as capabilities, authentication information, etc., to minimise the air interface signalling between the UE and the DBS when the HO is executed. Nonetheless, such techniques require a reliable and intelligent mobility prediction scheme tailored for the CDSA. To summarise Sections 2.2–2.4, Table 2.4 lists limitations of the conventional architecture along with the system improvements from the CDSA.
### 2.4. Centralised Control Plane

<table>
<thead>
<tr>
<th>Conventional architecture limitations</th>
<th>CDSA solutions and proposals</th>
<th>Scenarios of interest</th>
</tr>
</thead>
</table>
| High EC due to always-on service approach | Low EC with on-demand always-available system:  
  - DBS with on/off operation [16], [25], [77]  
  - DBS with high gain and selective beam-forming [26], [50] | Low utilisation, dense deployment |
| Wide area coverage may not be guaranteed at high frequency bands | Wide area coverage is provided at low frequency bands with dual connectivity:  
  - PCC: CBS and DBS at low and high frequency bands, respectively [6], [27], [49] | SC at high frequency bands |
| Scalability and coverage trade-off | DBSs can be gradually deployed when and where they are needed without coverage constraints:  
  - Cloud-HetNet: Radio resource heads with Cloud-RAN [50]  
  - Fusion network: on-demand deployment of DBSs [51] | SC at high frequency bands |
| Resource wastage with traffic tendency fluctuation due to (semi)static resource assignment constraints | Flexible DBS reconfiguration opportunities due to the absence of CRS:  
  - DBS with Dynamic TDD [54] | Dense deployment |
| Limited interference coordination between the cells based on local scope | Centralised interference coordination with a wide view of network status and parameters:  
  - CBS as a centralised coordinator [19] | Dense deployment |
| Resource selection is UE driven | Resource selection is network driven:  
  - Centralised RME for DBS-UE association [90] | Dense deployment |
| Poor mobility performance | UE is anchored to a MC with a large coverage area:  
  - Mobility performance depends on the macro layer [93], [96] | High user speed, dense deployment |


2.5 Challenges and Enabling Technologies

The CDSA aims to provide a framework where limitations of the conventional architecture can be overcome, as discussed in Sections 2.2–2.4. However, there are several research challenges and questions that need to be answered in order to concretely assess the feasibility and superiority of this architecture over the conventional one. These issues include: serving node selection, control message and data frame design, mobility and HO procedures, backhauling mechanisms, heterogeneous deployment with dual connectivity, channel estimation and management of discontinuous transmission techniques. In addition, the promotion of the CDSA as a candidate RAN for future cellular systems is tightly coupled to the emerging concept of SON [97]. This section discusses some of the CDSA challenges and provides a survey of the preliminary work that has been done to solve these issues. Furthermore, an overview of SON implementation in the CDSA is provided.

2.5.1 Context Information

Traditionally, the control and the data services are provided by the same physical node and the SS is usually used as a metric for serving node selection and HO decisions. Separating the CP from the DP makes these decisions non-trivial because different services (i.e., control and data) are provided by separate nodes. These nodes might be deployed at different locations and they could have different characteristics such as transmission power, antenna pattern, etc. Thus, the SS (at/from the CBS) cannot be used as a metric for selecting the serving DBS. Relying on measuring the DBS signal may not be feasible either because the best serving DBS may not be discoverable by the UE (e.g., switched off for energy saving or interference reduction). As a result, the DBS-UE association requires assistance by the CBS which brings several advantages as discussed earlier. However, such network driven approaches require intelligence, context awareness and CP/DP coordination.

Position information can be considered as the simplest metric in associating the UE with a DBS. Nonetheless, this criteria does not guarantee selecting the best serving DBS because obstacles and other loss components may exist in the path between the
UE and its nearest DBS. Broadly, radio channels between the UE and other DBSs may be better than the nearest DBS channel. Hence the CBS needs to obtain knowledge of channels conditions between the UE and each candidate DBS. Other parameters such as EC, mobility history, application requirements and network status can also affect the DBS selection decision [98]. For instance, assigning the UE to an already-awake DBS that is able to satisfy its requirements and excluding the inactive DBSs from the candidate set (where possible) could reduce the EC significantly. On the other hand, mobility pattern/history optimises the resource selection process for moving terminals. This highlights the importance of context awareness in the CDSA, which can be exploited to improve the EE, optimise the HO parameters and to design optimum traffic management policies.

Gathering the context information is one of the research challenges that need to be addressed. Some information can be easily and reliably gathered in current standards. For instance, position information can be provided by a global positioning system (GPS) or other mature techniques. However, new mechanisms are needed to predict channel conditions between the UE and each candidate DBS. In this area, [99] proposes a database-aided channel quality prediction technique for cellular systems with CP/DP separation. Each CBS is equipped with a database that contains SNR measurements for each DBS under its control. As shown in Fig. 2.8, this database maps each measurement to the geographical location where it is reported from, with Long. and Lat. being the location coordinates that depend on the required granularity, e.g., longitude and latitude respectively.

The authors of [99] use the SNR as a metric for channel quality prediction and argue that the SNR values consume less memory than the SINR. The former, i.e., the SNR, can be considered as a static measure that depends on the DBS power and antenna radiation pattern, the UE noise and the path loss model. Expressed differently, the SNR at a fixed location remains roughly the same as long as the serving DBS power and antenna pattern remain constant. On the other hand, the SINR depends on additional parameters such as network loading and number of interferers. For instance, the SINR at a fixed location at a certain time when all neighbouring DBSs are kept on and transmitting data will be lower the SINR at the same location at a different time when
2.5. Challenges and Enabling Technologies

some of the neighbouring DBSs are switched off. In other words, the instantaneous interference depends on network status. Thus storing SINR measurements of each state may not be feasible from a memory perspective although the SINR provides a better measure than the SNR.

The database training process requires the UE to report their locations along with the DBS pilot measurements to the CBS. If there is no previous measurement for the location being reported by the UE, the reported value is added to the database. Otherwise, an exponential moving average is used to incorporate the new value. In this way, the UE can measure pilot signals of the active DBSs and use the stored values of the inactive DBSs to determine the best channel quality. It can be noticed that this technique predicts the signal at the UE location given that there were previous measurements in this location. Nevertheless, it does not address the case when the propagation environment changes or when there is no previous reports from the UE position. This raises a question of how to deal with out-of-date measurements and how to interpolate between the database entries. Two approaches have been proposed in [100] to tackle these issues:

1. Historical SS fingerprint power map.

2. SS-to-distance map with inter-DBS measurements.
The first solution relies on a power map constructed by collecting SS fingerprints previously reported by UE at different locations. The expected SS at a specific UE position can be estimated by averaging the nearest fingerprints that are weighted according to their distance from the UE. This technique does not require channel modelling and it implicitly includes fading and non-line-of-sight effects. However, an accurate estimation of the SS requires a reliable fingerprint database, which can be obtained through time consuming drive tests. Wardriving (i.e., online database construction) and fingerprint prediction methods can also be used to construct the power map but they provide coarse and less accurate predictions [100].

The second approach relies on inter-DBS SS measurements to select the serving node. Given the location of each DBS, a SS-to-distance map is constructed, which can be used to estimate the expected signal at the UE location. In contrast to the power map technique, the SS-to-distance map method depends on measurements between the DBSs only. As a result, it does not require pre-configuration (i.e., drive tests) or propagation constants estimation. In addition, it can adapt to the propagation environment by updating the estimated signal according to the instantaneous inter-DBS measurements [100].

2.5.2 Self-Organising Networks

The prohibitive cost/effort for manual configuration and optimisation of network elements and parameters has motivated researchers and standardisation bodies to automate these procedures. In 3GPP parlance, such automatic operations are usually referred to as SON, which cover three areas [7]:

- **Self-configuration**: concerns with pre-operational procedures such as automatic configuration and integration of newly installed BSs in a plug-and-play mode.

- **Self-optimisation**: dynamically adjusts and optimises the operational characteristics in an automatic manner to cater for traffic patterns and propagation environment variations.
• Self-healing: minimises failure impact by identifying the failing element(s) and adjusting the appropriate parameters for service recovery.

As far as the CDSA is concerned, the self-optimisation capability can be considered as the most important aspect of SON. It can play a key role in enabling most of the CDSA applications especially those related to energy saving, load balancing and mobility robustness. With rapid traffic variations, switching the DBSs on/off and controlling their beam directions manually would not be feasible and may not be timely enough. A more convenient design approach is to automate these techniques in order to ease their implementation and to maximise their effectiveness. Such SON-based mechanisms allow recognition of short term energy saving opportunities and they enable proper reconfiguration of long term EE improvement strategies [101].

An energy self-optimising scheme has been proposed in [102] to automate the BS wakeup and hibernation process by using an online optimisation algorithm. Similarly, the 3GPP investigates several energy saving deployment scenarios with an underlay coverage layer provided by macro BSs of legacy networks. In these scenarios, RAN nodes of the capacity boosting layer are switched on/off automatically with commands issued by a centralised operation, administration and management entity or by the node itself according to certain criteria and polices [103].

By considering load balancing as the primary objective of the self-optimisation process, [104] proposes a mobility load balancing mechanism to shift cell edge users from congested to low utilised BSs whilst minimising the associated mobility overhead. Although this technique does not consider a CP/DP separation, the basic concept can be applied to the CDSA without significant modifications. A UE-like BS that uses a M2M interface to communicate with the UE has been proposed in [105] for load balancing purposes. This BS does not transmit any CRS and its M2M links with the UE are controlled by a macro BS in a master-slave relationship. This configuration can be considered as a SON-based CDSA, and the results reported in [105] show that this architecture maximises the throughput whilst balancing the network load.
2.6 Preliminary Standardisation Work

This section presents an overview of preliminary standardisation proposals related to the CDSA. In particular, it focuses on the aspects being studied in the 3GPP for future LTE releases and next generation cellular systems.

2.6.1 Dual Connectivity

In conventional HetNets, the independent operation of SC and MC layers raises several problems as discussed earlier. Thus the standardisation bodies are considering integrating these layers by allowing the UE to communicate simultaneously with both the SC and the MC. This is something referred to as dual connectivity and it is being investigated by the 3GPP for future LTE releases under the study item “small cell enhancements”. A very recent 3GPP technical report [106] indicates that the next generation RAN architecture will be designed to support this feature. According to [107] and [108], dual connectivity may imply CP/DP separation, UL/DL separation, RRC diversity or selective HO, as shown in Fig. 2.9.

- **Dual connectivity with CP/DP separation**: This is the classical scenario relevant to the CDSA, where data transmission and network connectivity are provided by different nodes, namely SCs and MCs, respectively, to minimise cell
planning effort and to improve mobility performance. As the user moves from one location to another, it transmits/receives data to/from the nearest SC. Mobility robustness is considered as one of the main benefits of dual connectivity with CP/DP separation [109]. However, new light-weight signalling and HO procedures are needed under this scenarios [109].

- **Dual connectivity with UL/DL separation:** The power imbalance between the MCs and the SCs in conventional HetNets implies that the best serving node in the DL may be different from the best one in the UL. Traditionally, the SS is used as a criteria for cell (re)selection, thus the high power of the MCs indicates that they would be better candidates (from a DL perspective) than the SCs even if the UE is in the vicinity of the latter. However, this may not be the case for the UL because the limited UE power suggests UL transmission to the nearest BS that offers the lowest path loss [110]. Cell-range extension can be used to increase the uptake area of low-power nodes, which are typically better UL (but not DL) choice. Expressed differently, current standards optimise either the UL or the DL performance. This trade-off can be avoided by offloading the UL traffic to the SCs whilst keeping the DL traffic in the MCs [109–111].

- **Dual connectivity with RRC diversity:** The dual connectivity is exploited in this case to provide the RRC signalling via multiple links in order to support a robust CP, as well as to enhance the mobility performance [108].

- **Dual connectivity with selective HO:** This scenario aims to provide different services via different nodes, e.g., high-rate best-effort services are provided by SCs while low-rate voice services are supported by MCs. This can be achieved by using different HO thresholds for different services [107].

The rest of this section focuses only on dual connectivity with CP/DP separation, since it is relevant to the CDSA. A CP/DP split model has been proposed in [112] based on protocol stack of the LTE. In this model, most of the data radio bearers (DRBs) are established at the SCs that have a direct interface with the serving gateway (S-GW). On the other hand, the signalling radio bearers (SRBs) and few DRBs are established at the
MCs that manage all the DRBs even though the latter are established at the SC layer. Only one RRC connection is established between the UE and the MC for connection control, while a new interface between the MC and the SC is used to exchange the critical and less dynamic information. Recently, [113] proposed procedures for the CBS to establish, modify and release the DBS DRBs. Two types of transmission/reception modes have been proposed in [114] to support the dual connectivity. These include simultaneous and time division multiplexing (TDM) modes. In the former, the UE transmits/receives data to/from both the SC and the MC at the same time. Although it utilises the resources efficiently, this mode may complicate the UE RF design and requires a careful power control when the MCs and the SCs are deployed in the same frequency band [114]. In other words, interference management can be considered as the main challenge of the simultaneous mode. On the other hand, the TDM mode segments the resources between the SC and the MC in the time dimension. At a given time (or subframe in LTE terminology), the UE transmits/receives to/from either the SC or the MC. In contrast to the simultaneous mode, the TDM mode does not require the UE to operate with two carriers at the same time, which relaxes the requirements on the UE RF capabilities [111]. However, it results into resource splitting loss and may require frequent resynchronisation when the UE switches from one carrier to the other. This suggests a careful design of the switching periodicity in the TDM mode to balance the resynchronisation and the data/control transmission delay. Fig. 2.10 shows the simultaneous and the TDM modes.

2.6.2 Lean Carrier

In conventional cellular systems, the CRS is transmitted in every subframe irrespective of the subframe activity status. In addition to ICI and SE loss, this periodic transmission of CRS prevents switching off the BS transmission circuitry during unoccupied subframes. Targeting a higher SE with minimum EC, [115] proposes a new carrier type, known as lean carrier, for future LTE releases. In the lean carrier, the CRS is replaced with a UE specific reference signal (UE-RS) for channel estimation along with a channel state information reference signal (CSI-RS) for channel quality measurements. Unlike the CRS, the UE-RS is transmitted only in resource blocks that contain users’ traffic
2.7 Conclusion

A comprehensive survey of RANs with CP/DP separation was presented. Potential benefits of this architecture and its superiority over the conventional one were critically discussed. In addition, the preliminary work to tackle its technical challenges was highlighted along with the ongoing discussion in standardisation forums related to this
2.7. Conclusion

research vision. Based on this survey, the following conclusions are drawn and potential research directions are underlined:

- Green communication is a hot topic nowadays. The mature and widely available EC profiles of the most power consuming element (i.e., the BS) highlight the importance of moving towards dynamic on/off BS operations. However, coverage constraints limit opportunities and gains of such techniques. This suggests a two layer RAN architecture with an underlay always-on connectivity layer complemented by an overlay on-demand data layer: the CDSA. In this area, there is a wealth of literature and proposals than investigates EE of the CDSA.

- Network densification has been accepted to be the dominant theme for future cellular systems. In such scenarios, the conventional RAN architecture, where each BS makes decisions based on its local scope only, may not be suitable. A more conscious approach will require centralised decision makers that have a global view of the network. The CDSA presents itself as a promising solution to address this issue by enabling the CBSs to act as centralised coordinators (and possibly decision makers) for the DBSs under their control. In this regard, the Cloud-RAN and the SDN concepts can play a key role.

- The evolution of current cellular generations is driven by performance improvement with an anywhere/anytime service paradigm. Nonetheless, futuristic demands, use cases and deployment scenarios require considering the signalling dimension to ensure efficient operation of cellular networks. In this direction, the conventional worst-case design approach may not be suitable and new adaptive signalling mechanisms are needed to minimise the overhead whilst maximising the data transmission resources.

- In the CDSA, the definition of coverage is different from the classical meaning that it has in conventional systems. Specifically, two types of coverage can be distinguished: area coverage and service coverage. A user with an area coverage (provided by the CBS) is a user that can camp on the network and issue a service request whenever needed. On the other hand, a user with a service coverage is
a user that can get the promised QoS such as the required data rate. Since a subset of the DBSs can be switched off during off-peak periods, some users (e.g., active and moving UE) may not get a service coverage although they have an area coverage. This suggests predictive mobility management where the UE trajectory is predicted in advance in order to optimise the service coverage whilst maximising the DBS sleep opportunities. Such techniques can also be used to improve HO performance whilst minimising the associated overhead and interruption time.

- The dual connection in the CDSA motivates a network design with a local mobility anchor at the CBS, resulting into a light-weight HO procedure between the DBSs. This approach can alleviate mobility overhead and minimise the CN signalling related to the HO process. Nonetheless, a research effort is needed to develop models for HO signalling cost/overhead under dual connection in order to define operational parameters that results into minimal CDSA HO signalling.

Based on this discussion, the next three chapters of this thesis focus on adaptive and predictive signalling for channel estimation and mobility management, respectively, under backhaul constraints. In particular, three main CDSA issues are addressed: adaptive DBS pilot signalling design and frame allocations to minimise the PL signalling overhead, predictive mobility management with two modes of operation to minimise the DBS HO latency and overhead, and CN-transparent HO signalling with backhaul latency constraints to minimise the CN signalling overhead.
Chapter 3

Correlation-based Adaptive DBS
Pilot Signalling

This chapter proposes a novel correlation-based adaptive pilot signalling pattern for the DL frame of the DBS under CDSA configuration. It shows that the proposed scheme can provide a significant saving in pilot signalling overhead without (or with a marginal) performance penalty as compared with the conventional worst-case design approach. The work presented in this chapter has been published in [8], [9] and [10].

3.1 Introduction

Meeting the ambitious 5G targets of 10 Gbps peak data rate and 1 ms roundtrip latency [1], [117] needs addressing a critical issue: signalling overhead. Current signalling mechanisms are designed to operate efficiently, to some extent, for current density levels. However, the dominant theme for future cellular systems, i.e., network densification, may not be suitable with these mechanisms due to the expected dramatical increase in signalling overhead.

Traditionally, all cellular users are connected to the same BS irrespective of their activity state (i.e., active, idle or detached) provided that they are within the footprint of this BS. Thus the same PL frame is used by all UE and hence most of the control
signals are cell-specific rather than user-specific resources. For example, the CRS of the LTE is used as a pilot by active and idle UE for channel quality measurements and for channel estimation to allow coherent demodulation of control and data channels. In addition, it is used in the initial access phase to demodulate the broadcast channel [118]. Since channel conditions of the detached and the idle UE are usually unavailable, these signals are distributed in the time/frequency grid based on the worst-case scenario, e.g., severe channel conditions and high mobility assumptions [119]. Although this approach guarantees acceptable performance for all users including those in severe conditions, it over-provisions the PL frame under moderate or good channel conditions.

In current standards, this overhead consumes a significant part of transmission resources. In the LTE for example, the CRS has a fixed overhead of 4.76% with one antenna port. To avoid CRS interference between different antenna ports of the same BS, the LTE adopts a shifted CRS pattern with resource muting [120]. As shown in Fig. 3.1, when an antenna port transmits its CRS, other ports mute their transmissions. Hence the CRS overhead increases to 14.25% with four antenna ports. Similarly, the cyclic prefix has a fixed overhead of 7.14% and 25% for normal and extended cyclic
3.1. Introduction

In the six middle resource blocks, four OFDM symbols in the second time slot of the first subframe are reserved for the broadcast channel, while the primary and the secondary synchronisation signals are transmitted every 5 ms and they occupy two OFDM symbols in each transmission [120]. The overall LTE frame overhead depends on the used configuration i.e., BW, antenna ports, duplex mode, etc. and it can reach up to 50% as shown in Fig. 3.2.

Some proposals to reduce this overhead are being considered such as using several classes of pilots with each class being transmitted at the necessary rate, e.g., high rate UE-RS for channel estimation and low rate CSI-RS for link adaptation measurements. Nonetheless, these signals also have a static pattern constrained by the worst-case conditions. In the CDSA, however, the DBS frame structure can be simplified and several signals can be removed. According to the mapping of Tables 2.1 and 2.2 of Chapter 2, the broadcast and the multicast bearer signals may not be required in the DP if these functionalities are delegated to the CP. Thus the DBS is invisible to both the idle and the detached UE and its on-demand connection with the active UE is established and assisted by the CBS. This relieves the DBS from the task of transmitting CRSs and removes the constraints imposed by the unknown channel conditions of the inactive UE.
As a result, the DBS-UE link lends itself to flexible, adaptive and optimised operations. For instance, the DBS pilot signal can be considered as a UE-specific resource and its transmission rate/pattern can be adaptively adjusted according to the temporal channel conditions between the DBS and the active UE. In this direction, an adaptive DBS pilot signalling pattern is developed in the following sections. It is worth mentioning that the contribution in this chapter does not provide a new mathematical model. Instead, the proposed adaptive scheme utilises existing models that have been used in other domains and adapts them to the considered pilot pattern estimation problem whilst considering the effects of noise and interpolation errors.

3.2 System Model

The system model considers a multi-carrier air interface for the DBSs. OFDM is adopted as an example because it has been accepted to be one of the strongest access technique candidates [1], [122]. However, the proposed scheme can be applied to other multi-carrier systems. The proposed method depends on the channel frequency response (CFR) correlation to (re)distribute the pilots dynamically according to channel variations. To minimise the feedback overhead, an equi-spaced pilot arrangement is considered which provides the optimal channel estimation performance [123].

The receiver (Rx) i.e., the active mobile terminal, calculates the correlation coefficients (CCs) between the estimated CFR and determines the maximum time/frequency spacing that provides a predefined correlation target. This result is fed back to the transmitter (Tx) i.e., the DBS, which redistributes the pilots accordingly. Fig. 3.3 shows system model of the proposed scheme. The pilot pattern estimation procedure is divided into two independent problems: FD and TD adaptive patterns. Combining these two together yields the optimal pilot arrangement for each UE. Subsection 3.2.1 develops the adaptive FD pilot signalling pattern and then the TD signalling pattern is determined in a similar manner in Subsection 3.2.2.
3.2. System Model

3.2.1 Adaptive Frequency Domain Pilot Signalling

The FD correlation function assuming perfect channel knowledge can be written as:

\[
R_H(\Delta_m) = \mathbb{E} \left[ H(m) H^*(m + \Delta_m) \right],
\]

where \( H(m) \) is the CFR at subcarrier \( m \), \( \mathbb{E} \) is the expectation operator, \( * \) is the complex conjugate and \( \Delta_m \) is the correlation lag. The normalised correlation function \( \hat{R}_H(\Delta_m) \) w.r.t. zero lag is:

\[
\hat{R}_H(\Delta_m) = \frac{\mathbb{E} \left[ H(m) H^*(m + \Delta_m) \right]}{\mathbb{E} \left[ |H(m)|^2 \right]} = \frac{R_H(\Delta_m)}{R_H(0)}.
\]

\( \hat{R}_H(\Delta_m) \) is inversely proportional to multipath delay spread which defines frequency selectivity of the channel. Conventional cellular systems consider the worst-case delay spread in dimensioning the FD pilot pattern irrespective of the actual channel conditions. In contrast, the proposed method estimates the actual CCs (and hence the channel frequency selectivity) which allows redistributing the pilots accordingly. Accurate estimation of the CFR is of great importance in determining \( \hat{R}_H(\Delta_m) \). Both estimation and interpolation errors affect the estimated CCs and the pilot pattern. A
simple least square (LS) estimator is considered and the following derivations investigate and eliminate the noise effects on the correlation function.

In a typical OFDM Rx, the LS estimate of the channel at subcarrier \( m \) can be expressed as [124]:

\[
\tilde{H}(m) = \frac{Y(m)}{X(m)} = H(m) + Z(m),
\]

(3.3)

where \( Y(m) \) and \( X(m) \) are the received and the transmitted signals, respectively, at subcarrier \( m \) in the FD. \( Z(m) = \frac{N(m)}{X(m)} \) is the channel estimation error with \( N(m) \) being the \( m \)th subcarrier FD noise component which is modelled as an independent and identically distributed additive white Gaussian noise (AWGN) with zero mean. Replacing \( H(m) \) in (3.1) with \( \tilde{H}(m) \) in (3.3) gives the CCs based on real i.e., noisy, channel estimates.

\[
R_{\tilde{H}}(\Delta m) = E[H(m)H^*(m + \Delta m)] + E[H(m)Z^*(m + \Delta m)] + E[Z(m)H^*(m + \Delta m)] + E[Z(m)Z^*(m + \Delta m)].
\]

(3.4)

The AWGN is random and uncorrelated with the channel and the transmitted signal i.e., \( H \) and \( Z \) are uncorrelated. Thus \( E[H(m)Z^*(m + \Delta m)] = E[Z(m)H^*(m + \Delta m)] = 0 \), and (3.4) simplifies to

\[
R_{\tilde{H}}(\Delta m) = E[H(m)H^*(m + \Delta m)] + E[Z(m)Z^*(m + \Delta m)]
\]

\[
= R_H(\Delta m) + E[Z(m)Z^*(m + \Delta m)].
\]

(3.5)

Equation (3.5) indicates that the correlation function based on a noisy LS channel estimate consists of the ideal CCs in addition to a noise bias. This effect can be seen in the simulated correlation function of Fig. 3.4. Since the AWGN is random and uncorrelated (i.e., the correlation between AWGN components at different subcarriers is zero), the term \( E[Z(m)Z^*(m + \Delta m)] \) simplifies to 0 when \( \Delta m \neq 0 \). Notice that when oversampling and/or pulse shape filtering are used at the Rx side, then the noise after the filtering/oversampling operation may or may not be correlated depending on the autocorrelation function of the filter [125]. However, this case has not been considered in this thesis. When \( \Delta m = 0 \), the term \( E[Z(m)Z^*(m + \Delta m)] \) becomes an autocorrelation which gives the channel estimation error variance. As can be seen in Fig. 3.4a, the correlator removes the bias introduced by the noise at all other lags except
3.2. System Model

![Graph](image)

(a) Absolute correlation function

![Graph](image)

(b) Normalised correlation function

Figure 3.4: Noise effects on the estimated correlation function. SNR = 3 dB

the zero lag because \( Z(m) \) and \( Z(m + \Delta_m) \) are uncorrelated when \( \Delta_m \neq 0 \). However, normalising the estimated (and unbiased) CCs w.r.t. the biased zero lag biases the correlation function as depicted in Fig. 3.4b.

Since the CCs at all other lags except the zero lag do not suffer from a noise bias, they can be used to predict the zero lag CC by means of polynomial fitting. This approach has been proposed in [126] to remove the AWGN effect from a time domain covariance-based velocity estimator, and a similar method has been adopted in [127]. Based on the CCs at all other lags except the zero lag, a polynomial can be formulated
as:

\[ R_H(\Delta_m) \approx \sum_{i=0}^{\hat{c}} \hat{a}_i (\Delta_m + \hat{b})^i, \]  

(3.6)

where \( \hat{c} \) is the polynomial order, \( \hat{a}_i \) is the \( i^{th} \) polynomial coefficient and \( \hat{b} \) is the base of the correlation function. Assuming that the first subcarrier is used as a base (i.e., \( \hat{b} = 1 \)), the normalised noise-compensated FD correlation function can be obtained by dividing the estimated correlation function, i.e., (3.4) or (3.5), by the approximated zero lage CC from (3.6), i.e.,

\[ \hat{R}_H(\Delta_m) = \frac{R_H(\Delta_m)}{\sum_{i=0}^{\hat{c}} \hat{a}_i}. \]  

(3.7)

The model presented so far investigates and eliminates the channel estimation errors from the correlation function. However, channel estimation is performed at pilot subcarriers only and hence an interpolator is needed to estimate the channel at data subcarriers. Thus, (3.7) can be affected by interpolation errors especially at low pilot densities (i.e., large spacing between the pilots). To overcome this problem, one of the radio frame\(^1\) symbols is used as a training symbol for all users to obtain their correlation functions. This eliminates the need for a FD interpolator at the training symbol, hence the FD CCs are not affected by any interpolation errors.

**Training Symbol Validation**

Assuming that the delay spread experienced by each user is fixed over the duration of one radio frame but may change from a radio frame to another, different users scheduled at different subframes within the same radio frame can obtain the FD CCs based on the estimated CFR at the same training symbol. Then, the result can be used as an indicator for the correlation function at each user’s subframe. This assumption can be validated by considering the separation property of wireless channels. Given the CFR at the \( m^{th} \) subcarrier and the \( n^{th} \) OFDM symbol:

\[ H(m, n) = \sum_i \gamma_i(nT) e^{-j2\pi mi/M}, \]  

(3.8)

\(^1\)Aggregate of several subframes
where $\gamma_i$ is the gain of the $i^{th}$ path. $T$ is the OFDM symbol duration and $M$ is the total number of subcarriers. As shown in [128], the joint frequency-time correlation function of the CFR $r_H(\Delta_m, \Delta_n)$ can be decoupled into a multiplication of a frequency correlation $r_H(\Delta_m)$ and a time correlation $r_H(\Delta_n)$, i.e.,

$$r_H(\Delta_m, \Delta_n) = P_H^2 r_H(\Delta_m) r_H(\Delta_n), \quad (3.9)$$

with

$$r_H(\Delta_m) = \sum_i \left( \frac{P_i}{P_H} \right)^2 e^{-j2\pi \Delta_m i/M}, \quad (3.10)$$

$$r_H(\Delta_n) = J_0(2\pi D_s T \Delta_n), \quad (3.11)$$

where $P_H^2 = \sum_i P_i^2 = 1$ is the total average power of the normalised channel impulse response, $J_0$ is the zero order Bessel function of the first kind, $D_s$ is the Doppler shift.

The reader is referred to [128] for the derivation of these equations. It can be noticed that (3.10) does not depend on the symbol index (i.e., $n$) or the time separation (i.e., $\Delta_n$). Thus for each user, all OFDM symbols of the radio frame have the same FD correlation function irrespective of the time variations which validates the approach of using one training symbol to obtain the FD CCs regardless of the UE’s allocations within the radio frame.

From an overhead perspective, the training symbol has a negligible impact because it consumes one OFDM symbol only across the whole radio frame. For example, considering a typical LTE radio frame with 10 subframes each one consisting of 14 symbols, the training symbol overhead is only 0.7%. In addition, the training symbol can be the same symbol used for synchronisation or it can be used as a pilot for users allocated to the subframe where this symbol is transmitted. In the latter case, the training does not require dedicated resources and hence it will not introduce any additional overhead.

Finally, each Rx can determine the maximum allowed frequency spacing between the pilots ($\tilde{N}_f$) for a certain FD correlation value ($\Upsilon_f$) as:

$$\tilde{N}_f = \max_{R_H(\Delta_m) \geq \Upsilon_f} \Delta_m. \quad (3.12)$$

For an ideal exponential power delay profile (PDP), the theoretical FD pilot spacing
can be formulated as [129]:

$$N_f = \frac{\sqrt{\left(\frac{1}{\tau_f}\right)^2 - 1}}{2\pi \tau_{rms} \delta_f},$$  \hspace{1cm} (3.13)

where $\tau_{rms}$ is the rms delay spread and $\delta_f$ is the subcarrier spacing. Equation (3.13) is included herein as a reference for evaluations only.

### 3.2.2 Adaptive Time Domain Pilot Signalling

Following a similar approach, the pilot signalling pattern in the TD can be adaptively adjusted according to the TD CC. The latter is affected by Doppler shift which itself depends on the user’s speed. For an accurate estimation of the TD CCs, one subcarrier across the entire time/frequency resource grid will be used as a training subcarrier. By exploiting the separation property, the TD CCs based on the training subcarrier can be used for all other subcarriers irrespective of the frequency variations. Thus different UE can obtain their TD CCs by using one subcarrier only which introduces a negligible overhead. For instance, using one of the 1200 data subcarriers for the training purpose results into 0.08% overhead in a typical LTE system with 20 MHz BW.

The maximum allowed time spacing between the pilots ($\tilde{N}_t$) for a certain TD correlation value ($\Upsilon_t$) can be formulated as:

$$\tilde{N}_t = \max_{\hat{R}_H(\Delta_n) \geq \Upsilon_t} \Delta_n,$$  \hspace{1cm} (3.14)

where $\hat{R}_H(\Delta_n)$ is the normalised TD CCs which can be calculated in a similar way as $\hat{R}_H(\Delta_m)$. Fig. 3.5 shows example allocations of a DBS DL resource grid with the proposed adaptive pilot signalling pattern. Notice that the pilot allocations can be uniform across the entire resource grid (i.e., according to the worst-case conditions among the active UE), or each active UE can have a tailored pilot pattern as shown in Fig. 3.5. However, the latter case may require a coordination between neighbouring DBSs and/or between antenna ports of the same DBS to avoid pilot interference.

It is worth mentioning that under good channel conditions e.g., very small delay spread or very low speed, $\tilde{N}_f$ or $\tilde{N}_t$ respectively can result into a very low pilot density, e.g., one pilot. In this case, the estimated channel at this pilot reflects the stationery
channel at data resources. However the equalised data symbols could be subject to a noise depending on the used estimator and equaliser. This suggests restricting a minimum pilot density in the FD and the TD for noise averaging. It can be noticed that the adaptive scheme introduces an additional signalling overhead in the UL because each UE estimates the optimal pilot pattern and sends the results to the serving DBS. Nonetheless, this additional UL overhead is marginal because only the pilot spacing needs to be reported. For instance, considering the typical LTE subframe with 14 OFDM symbols, only 4 bits are required to report a maximum pilot spacing of 14 symbols in the TD. A similar number of bits will also be required to report the FD pilot spacing.

### 3.3 Performance Analysis

Link level simulations have been performed to assess performance and potential overhead savings of the proposed adaptive mechanism. First, the accuracy of the correlation-based pilot signalling pattern estimation is compared with the theoretical spacing of an exponential PDP. The comparison is based on the normalised mean square error
3.3. Performance Analysis

Figure 3.6: Pilot signalling pattern estimation accuracy and potential overhead saving of the proposed adaptive scheme. Simulation parameters: exponential PDP, $\Upsilon_f = \Upsilon_t = 90\%$, FFT size = 2048, used subcarriers = 1200, OFDM symbols = 140, minimum pilot density = 2.

(NMSE) given by (3.15) and the results are reported in Fig. 3.6a.

\[
\text{NMSE} = \frac{\mathbb{E} \left[ (N_f - \tilde{N}_f)^2 \right]}{\mathbb{E} \left[ (N_f)^2 \right]}.
\]

As shown in Fig. 3.6a, the adaptive scheme estimates the required spacing between the pilots with NMSE $\approx 4\%$, which is roughly the same NMSE of an ideal scenario.
where the Rx has a perfect channel knowledge. In addition, the proposed method is almost insensitive to the noise level, i.e., it can compensate the noise bias at the zero lag perfectly, hence it performs in low SNR conditions as good as in high SNR ranges. In other words, the adaptive scheme provides a reliable estimation of the required pilot spacing irrespective of the AWGN level. In order to evaluate potential gains of the proposed mechanism, the theoretical overhead of the adaptive pattern is compared with pilot pattern of the worst-case: 500 km/hr speed and 991 ns rms delay spread. As can be seen in Fig. 3.6b, the adaptive scheme could significantly reduce the pilot signalling overhead. Considering a typical LA scenario with $\tau_{rms} = 100$ ns [130] and 30 km/hr speed as an example, the proposed scheme could reduce the pilot signalling overhead by more than 90% w.r.t. the static worst-case pattern. As expected, the potential gains (i.e., in terms of overhead saving) decrease as the speed and/or the delay spread increase.

A second set of simulations is performed to determine gains and performance penalty of the adaptive pilot scheme. A typical LTE CRS pattern for one antenna port [120] is considered as the static/worst-case dimensioned pattern. For simplicity, a full buffer traffic model and a single modulation/coding scheme are adopted. The developed simulator assumes that the users are multiplexed in a time division scheme. In LTE terminology, this means that different users do not share the same subframe, i.e., all resource block pairs of a given subframe are allocated to a single user only. Thus the adaptive FD pilot pattern expands the whole range of used subcarriers while the adaptive TD pilot pattern is bounded by the subframe duration. The minimum pilot density in the FD and the TD is set to two and other simulation parameters are provided in Table 3.1.

Fig. 3.7a compares the bit error rate (BER) of the adaptive scheme with the static/worst-case dimensioned pattern (i.e., the LTE CRS) while Fig. 3.7b shows the average pilot signalling overhead reduction in the considered simulation scenario. The training resources overhead (i.e., 0.7% for the FD training symbol and 0.08% for the TD training subcarrier) is accounted for whilst calculating the overhead of the adaptive scheme. It can be noticed that the adaptive pattern provides roughly the same BER performance as the worst-case dimensioned pattern with a slight degradation at high
Table 3.1: Link level simulation parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modulation</td>
<td>QPSK</td>
<td>Subcarrier spacing</td>
<td>15 kHz</td>
</tr>
<tr>
<td>Coding</td>
<td>Turbo</td>
<td>Cyclic prefix</td>
<td>Normal LTE cyclic prefix</td>
</tr>
<tr>
<td>Estimator</td>
<td>Least square</td>
<td>Channel model</td>
<td>Extended Pedestrian-A (EPA)</td>
</tr>
<tr>
<td>Interpolator</td>
<td>Linear</td>
<td>Doppler shift</td>
<td>5 Hz</td>
</tr>
<tr>
<td>Equaliser</td>
<td>Zero forcing</td>
<td>Total bandwidth</td>
<td>20 MHz</td>
</tr>
<tr>
<td>FFT size</td>
<td>2048</td>
<td>Guard band</td>
<td>10%</td>
</tr>
<tr>
<td>Used subcarriers</td>
<td>1200</td>
<td>Frequency</td>
<td>2 GHz</td>
</tr>
</tbody>
</table>

SNR values. As can be seen in Fig. 3.7a, increasing the correlation target improves the performance because a small spacing between the pilots is required to achieve such high correlation. With a 95% correlation target for the adaptive system, the performance is almost the same as the worst-case pattern for SNR values less than 10 dB. However at SNR $\geq 10$ dB, the adaptive pattern degrades the performance by $0.7 \text{--} 1.6$ dB. This can be traced to the fact that noise effects are marginal in the high SNR range, hence the interpolation errors dominate. Since the proposed scheme adapts the pilot spacing according to the channel variations, the pilot interval would be larger in the adaptive system as compared with the worst-case dimensioned system. Thus, at a high SNR, the interpolation errors have a higher impact on the performance of the adaptive pattern. Nonetheless, using robust interpolation techniques such as low pass or fast Fourier transform (FFT) interpolators could eliminate this difference.

Compared with the LTE CRS pattern, Fig. 3.7b indicates that the proposed mechanism reduces the DBS pilot signalling overhead by 74 -- 78% in the considered simulation scenario. In addition, the overhead gain increases as the correlation target decreases. Expressed differently, there is a trade-off between the performance and the overhead that can be controlled by the FD and the TD correlation values, i.e., $\Upsilon_f$ and $\Upsilon_t$ respectively.
Figure 3.7: BER performance and overhead reduction in the proposed adaptive scheme compared with the LTE CRS pattern

3.4 Conclusion

This chapter investigated one of the CDSA PL benefits in the signalling dimension. It has been shown that the conventional worst-case design approach over-provisions the PL frame, which suggests adaptive frame allocation techniques. However, the applicability of such techniques is very limited in the conventional RAN architecture due to the tight coupling between network access and data transmission services. On
the other hand, the CDSA frame structure offers more flexibility due to the decoupled nature of coverage and data services.

Since DBSs of the CDSA are invisible to both idle and detached UE, only channel conditions of active UE need to be taken into account whilst dimensioning the DBS PL frame. In other words, the feature of the DBS as a serving node for active UE only lends the DBS-UE link to flexible operations and dynamic allocations rather than adopting the conventional worst-case design approach. In this direction, an adaptive DBS pilot signalling scheme has been proposed. The developed model considers both the frequency and the time variations, and it depends on estimating the actual channel correlation to dynamically redistribute the pilot signals.

Simulation results depict that the proposed scheme can significantly reduce the pilot signalling overhead as compared with the worst-case design approach. In the considered simulation scenarios, up to 78% reduction in pilot signalling overhead is achieved without (or with a marginal) performance penalty. It has also been shown that there is an overhead/performance trade-off that can be controlled by tuning the correlation target. Although the proposed adaptive scheme considers channel estimation pilots only, the basic concept can be extended to the cyclic prefix (in OFDM systems) and other signals in order to minimise the overall DBS PL signalling overhead.
Chapter 4

Predictive DBS-Level Handover Signalling

This chapter proposes two novel predictive DBS HO schemes with advance DBS signalling for HO preparation and resource reservation. These predictive schemes aim to minimise the DBS HO signalling latency rather than minimising the HO signalling overhead. The work presented in this chapter has been published in [11], [12] and [13].

4.1 Introduction

The discussion in Chapter 2 draws a key conclusion on the importance of network densification in the future cellular system. With ultra-dense SC deployments, mobility management becomes complex due to the expected increase in HO rates. In such scenarios, the conventional HO procedures may lead to a dramatical increase in signalling overhead. For instance, the results reported in [6], [92] and [94] indicate high signalling overhead and call drop rates when the conventional HO mechanisms are applied in dense SC deployment scenarios. This suggests a paradigm shift towards a signalling conscious cellular architecture with intelligent mobility management.

A revisit to the conventional HO scheme is necessary to underline potential signalling improvement mechanisms. Without loss of generalisation, each conventional HO generates three types of signalling:
4.1. Introduction

- Air interface signalling.
- RAN signalling.
- CN signalling.

The air interface signalling includes measurement reports (MRs) that are reported, either periodically or on an event basis, to the serving BS. These reports provide information on SS and/or signal quality (SQ) 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 CN signalling. An example of this procedure is provided in Section 4.2.

In the CDSA, the centralised CP could offer simple and robust HO procedures with minimal CN signalling overhead. This can be utilised by exploiting the CBS as an RRC anchor point. Such an approach, here referred to as CN-transparent HO, is proposed and modelled in Chapter 5. From the RAN signalling perspective, however, the pure CDSA model may provide marginal gains since a DP HO is always required when the UE moves from one DBS to another. In addition, the green CDSA implementation with dynamic on/off DBS operation discussed in Section 2.3 can create “DP holes” for moving terminals. When an idle UE requests a service, the CBS selects the best candidate DBS and turns it on if necessary. However, some of the active UE may arrive at or pass by an inactive DBS and hence they may loose the DP connection. Context information and mobility prediction can play a key role in solving these problems, i.e., the HO-related RAN signalling and the DP holes in the CDSA. A reliable prediction of the user’s trajectory (at cell level rather than the exact location) allows the candidate DBS in the UE path to prepare and reserve resources in advance, which in turn could simplify the HO process and minimise the associated RAN signalling and interruption time. In addition, it allows the CBS to determine the candidate DBSs in the user’s path in order to send them an activation signal before the user reaches these
DBSs. Such a prediction can also minimise the unnecessary HOs and it may reduce the scanning/monitoring load on the UE.

In the conventional RAN architecture, the predictive strategies have tight constraints since an incorrect prediction with a break-before-make HO can lead to detaching the UE from the network. In other words, an incorrect prediction in the conventional RAN does not only increase the HO latency and signalling overhead, but also it requires a new UE-network connection establishment. On the other hand, the CDSA offers relaxed constraints in implementing predictive HO management strategies. An incorrect DBS prediction in the CDSA does not require UE-network connection re-establishment, since the UE maintains another low rate connection with the CBS. In this direction, predictive mobility management at DBS-level under CDSA is proposed in this chapter. With the main objective of minimising the DBS HO latency, two predictive HO schemes are proposed:

- History-based predictive DBS HO: it depends on mobility history to predict future DBS HO events. Based on a Markov Chain modelling, this scheme uses an online learning process to predict users’ trajectory in terms of a DBS HO sequence, which gives the UE trajectory at DBS level rather than the exact UE location. 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.

- Measurement-based context-aided predictive DBS HO: it exploits measurement trends coupled with UE context (i.e., direction, speed and history) to predict future DBS HOs along with the expected HO time. Grey system theory is adopted as a trending mechanism to predict the UE-DBS measurements in advance. In addition, the UE direction and speed are utilised to minimise processing and storage requirements, while the UE HO history is used as an added measure to confirm the predicted DBS. Furthermore, a UE-specific triggering threshold is proposed to improve the prediction performance whilst minimising the UE processing requirements.

In both schemes, 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/latency. Moreover, both schemes include 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 remainder of this chapter is structured as follows: Section 4.2 describes the HO signalling procedure with and without mobility prediction and formulates the HO latency in each scenario. Section 4.3 develops the history-based predictive DBS HO scheme and formulates the learning and prediction processes. Section 4.4 develops the measurement-based context-aided predictive DBS HO scheme and describes the operation of each unit along with their interactions. Section 4.5 provides simulation and numerical results, and discusses the impact of several parameters on the prediction and the HO performance. In addition, it shows gains of the proposed schemes over the conventional HO mechanism in terms of reduction in HO signalling latency. Finally, Section 4.6 concludes this chapter.

4.2 Handover Signalling Procedure and Latency

As mentioned in Section 4.1, predictive mobility management could allow light-weight and signalling-efficient HO procedures with minimal latency and overhead. To support this claim without loss of generalisation, the typical LTE X2 HO procedure is considered as an example for the conventional non-predictive HO scenario. In the latter, the UE measures signals of the detectable DBSs and reports the result to the serving (i.e., the source) DBS whenever the HO criteria is satisfied. The conventional HO procedure consists of three major steps:

- Preparation.
- Execution.
4.2. Handover Signalling Procedure and Latency

- 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 and updates the UE tracking area [131].

In the predictive HO procedure, most of the HO preparation steps can be completed before the HO criteria is met, provided that the prediction outcome satisfies the predictive HO decision criteria. 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 MR 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. 4.1 shows the signalling flow diagram for these cases.

The HO signalling latency can be expressed in terms of the delay required to transmit and process the HO messages [132], [133]. Denote $\zeta_{i,j}$ as the one way transmission delay from node $i$ to node $j$, $\varepsilon_j$ as the processing delay in node $j$. The HO signalling latency $\mathcal{H}$ can be written as:

$$\mathcal{H} = \sum \zeta_{i,j} + \sum \varepsilon_j,$$

(4.1)

where the summation in (4.1) is for all the nodes involved in the signalling flow of the HO process. In other words, each signalling message increases $\mathcal{H}$ by a transmission delay $\zeta_{i,j}$ and a processing delay $\varepsilon_j$. Since the actual HO procedure starts after the
source DBS receives the MR, the HO signalling latency includes the HO decision and the subsequent steps (depending on the HO type). Expressed differently, the advance preparation procedure (i.e., steps a, b and c of Fig. 4.1) is not included in the latency
4.3 History-based Predictive DBS Handover

This scheme relies on representing the DP network (i.e., the DBSs) 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 [134]. 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. 4.2 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 [134]:

$$p_k = p_0 T^k,$$

with

$$p_k = [p_1 \ p_2 \ p_3 \ \ldots \ \ p_n],$$

$$p_0 = [b_1 \ b_2 \ b_3 \ \ldots \ \ b_n],$$

Figure 4.2: Discrete-time Markov Chain with $n$ states (i.e., DBSs), only states 1, 2 and $n$ are shown for readability
where \( \mathbf{p}_k \) is the \( k \)th HO probability vector, i.e., \( p_i \) is the probability of being at DBS\(_i\) after \( k \) HOs. \( \mathbf{T} \) is the transition probability matrix while \( \mathbf{p}_0 \) is the initial distribution vector with \( b_i = 1 \) if the user starts the movement at DBS\(_i\) and 0 otherwise. Equation (4.2) 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, the learning procedure for updating the transition matrix is described.

### 4.3.1 Transition Matrix: Properties and Conditions

Consider \( \mathcal{M} \) as the DTMC state space with \( \mathbb{I} \) being the states’ indices. Define \( \mathcal{N}_i \) as a list of the DBSs that are neighbours\(^1\) to DBS\(_i\) \( \forall i \in \mathbb{I} \). Notice that \( \mathcal{N}_i \) 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:

\[
\mathcal{M} = \{ \text{DBS}_1, \text{DBS}_2, ..., \text{DBS}_\hat{n} \} , \; \mathbb{I} = \{ 1, 2, ..., \hat{n} \} ,
\]

where \( \hat{n} \) is the prediction set size. For DP HO prediction, \( \hat{n} \) represents the number of DBSs per CBS.

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

\[
0 \leq t_{i,j} \leq 1 , \; \forall i, j \in \mathbb{I}.
\]

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

\[
t_{i,i} = 0 , \; \forall i \in \mathbb{I}.
\]

\(^1\)The first tier neighbours that can be reached directly in a single HO.
• The direct HOs are possible between neighbouring DBSs only:

\[ t_{i,j} = t_{j,i} = 0 \quad \forall j \notin N_i. \quad (4.6) \]

• Any new movement starts from the destination of the previous trajectory. Thus the UE will definitely make an outbound\(^2\) HO from any DBS. However, the UE may not necessarily perform an inbound HO to all the DBSs in the network. As a result, \( T \) is a right stochastic matrix (a matrix where the summation of each row is one) but not necessarily a doubly stochastic matrix (a matrix where the summation of each row is one and the summation of each column is one). This property sets the following condition:

\[ \sum_{j=1}^{\hat{n}} t_{i,j} = 1 \quad \forall i \in \mathbb{I}. \quad (4.7) \]

### 4.3.2 Transition Matrix Initialisation

For each user, a \( \hat{n} \times \hat{n} \) transition matrix is constructed and initialised according to the conditions of Section 4.3.1. The process of initialising \( T \) involves invoking conditions (4.5) and (4.6) 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 \( T \) 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.

**Algorithm 1 Initialisation of the transition matrix**

1: Invoke conditions (4.5) and (4.6).

2: Set \( t_{i,j} = 1 \quad \forall j \in N_i. \)

3: Set \( t_{i,j} = \frac{t_{i,j}}{\sum_{j=1}^{\hat{n}} t_{i,j}} \quad \forall i, j \in \mathbb{I}. \)

\(^2\)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.
4.3.3 Online Learning Process

The transition matrix can be updated based on each UE HO history. However, maintaining the HO history/frequency for each UE may not be feasible from a memory perspective and could lead to a dramatical increase in storage overhead especially in dense deployment scenarios. In order to reduce the HO latency and improve the performance with minimal storage requirements, an online learning process is proposed. Fig. 4.3 shows a block diagram of the proposed history-based DBS HO learning and prediction scheme. 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 $\epsilon$, where $0 \leq \epsilon \leq 1$, is proposed to control the model’s reaction to random or less frequent movements. Small (large) values of $\epsilon$ 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 $T$. The extreme case of $\epsilon = 0$ means that $T$ will not be updated (hence the prediction is independent of the movement history), while the case of $\epsilon = 1$ biases the prediction towards the most recent trajectory.

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

\[ \mathcal{P}_1 = \{\text{DBS}_j : j \in N_v \land t_{v,j} > 0\} \cup \{\text{DBS}_w\}, \]

where \( \mathcal{P}_i \) is the players set in stage \( i \geq 1 \) of the game. In the first stage, the probability of the direct HO from DBS\textsubscript{v} towards DBS\textsubscript{w} is increased by a certain amount controlled by \( \epsilon \). Similarly, the probabilities of the direct HOs from DBS\textsubscript{v} towards all other playing DBSs (i.e., except DBS\textsubscript{w}) are decreased. This can be expressed mathematically as:

\[
t_{v,w}^{(1)} = t_{v,w} + \sum_j t_{v,j} \epsilon, \quad \forall \text{DBS}_j \in \mathcal{P}_1 \setminus \{\text{DBS}_w\}, \tag{4.9}
\]

\[
t_{v,j}^{(1)} = t_{v,j} - \frac{\sum_j t_{v,j} \epsilon}{\#\mathcal{P}_1 - 1}, \quad \forall \text{DBS}_j \in \mathcal{P}_1 \setminus \{\text{DBS}_w\}, \tag{4.10}
\]

where \( \#\mathcal{P}_i \) is the cardinality of the set \( \mathcal{P}_i \), the superscript \( (i) \) means the probability after stage \( i \). It can be noticed in (4.9) and (4.10) that larger \( \epsilon \) values result into an aggressive increase and decrease in \( t_{v,w}^{(1)} \) and \( t_{v,j}^{(1)} \), respectively. This suggests setting larger \( \epsilon \) values for users with highly regular mobility profiles. On the other hand, a smaller \( \epsilon \) setting may be appropriate for users with random mobility profiles. Such users add a higher noise to the prediction process due to the random nature of their movements. Hence a smaller \( \epsilon \) setting for random users improves the prediction performance because each HO results into a small change in the probabilities when \( \epsilon \) is small. The effect of this parameter on the prediction performance and the HO latency is discussed in details in Section 4.5.1.

It can be noticed that the first stage may violate condition (4.4) because \( t_{v,w} \) and \( t_{v,j} \) 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 \( T \). However this may lead to violating condition (4.7) 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 4.3.1 or affecting the learned history. In stage \( i > 1 \), the DBSs with zero or
negative probabilities after stage $i - 1$ leave the game. The DBSs with positive probabilities are called survivals and they equally share the negative probabilities resulted from stage $i - 1$. In other words, the player set in stage $i > 1$ includes the survivals only, i.e.,

$$\mathcal{PS}_i = \left\{ \text{DBS}_j : t^{(i-1)}_{v,j} > 0 \land \text{DBS}_j \in \mathcal{PS}_{i-1} \right\}. \quad (4.11)$$

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

$$t^{(i)}_{v,j} = t^{(i-1)}_{v,j} + \sum_{n \in \mathcal{PS}_{i-1}} \frac{t^{(i-1)}_{v,n}}{\#\mathcal{PS}_i}, \quad \forall \text{DBS}_j \in \mathcal{PS}_i, \quad (4.12)$$

where $i > 1$, $\text{DBS}_n \in \mathcal{PS}_{i-1}$ and $t^{(i-1)}_{v,n} < 0$. Notice that the second term of (4.12) is negative (i.e., the summation in (4.12) is for the negative probabilities that resulted from stage $i - 1$). 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 (4.2). 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 (4.2) with $k = 1, 2, 3, \ldots$ and $b_i = 1$ for the source DBS, and then selecting the DBS with the highest probability in each HO i.e.,

$$k^{th} \text{ HO DBS} = \text{DBS}_p \bigg|_{p_k = \max(p_k)}. \quad (4.13)$$

It is worth mentioning that the history-based predictive HO 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 a HO mode selection unit is proposed, 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 [135]. For the history-based predictive DBS 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.,
\[ h(p_k) = - \sum_{i=1}^{n} p_i \log p_i, \]  
(4.14)

where \( h(p_k) \) is the entropy of the district probability distribution \( p_k \) given by (4.2). The entropy unit 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 followed if there is a high confidence in the predicted target DBS, i.e., \( h(p_k) \leq h_{thr} \). On the other hand, the conventional non-predictive HO procedure is followed if the entropy of the predicted target DBS does not satisfy the confidence threshold, i.e., \( h(p_p) > h_{thr} \).

Based on this model, the predicted DBS, i.e., \( DBS_p \) in (4.13) with \( h(p_k) \leq h_{thr} \), can be sent an activation message before the UE reaches this DBS. In addition, the HO-related resource reservation and RAN signalling can be performed in advance to minimise the HO latency. However, this model predicts the target DBS without providing an indication of the expected HO time. The latter is an important parameter that 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. In order to include the time dimension in the prediction outcome, the following section develops a measurement-based context-aided predictive DBS HO scheme that includes the history-based scheme as one of the prediction components.

### 4.4 Measurement-based Context-aided Predictive DBS Handover

#### 4.4.1 System Model

This scheme depends on SS and/or SQ prediction performed by the UE. This prediction is aided by UE context information such as location, direction and speed, in addition
to statistical historical information based on either the UE HO history or the aggregated per-DBS Neighbour List (NL) HO history. The predicted DBS is reported on an event basis to the serving (i.e., the source) DBS, which decides the reporting criteria and the HO mode to be followed by the UE. The prediction process is triggered only once when a certain prediction triggering threshold is reached. As shown in Fig. 4.4, the measurement-based context-aided predictive DBS HO scheme comprises a location and span estimation unit, a SS and SQ prediction unit, a history prediction unit, a prediction analysis unit, a reporting unit, a HO mode decision unit and a HO mode switch unit.

The UE periodically measures SS and SQ of the serving DBS and the top-$\bar{n}$ other detectable DBSs at every measurement gap as in current standards. The 3GPP Measurement Reporting and Control (MRC) [136] may be re-used as an example of this measurement and optional reporting mechanism. The reported strongest or best quality DBSs are limited to $\bar{n}$ per DBS categorisation as the measurement interval is limited and measurement power consumption and normal transmission need to be balanced against the accurate MR cycle. The SS and SQ prediction unit stores measurements of a subset of the top-$\bar{n}$ detectable DBSs that reside within the angular span of the UE direction/speed. The location and span estimation unit triggers the prediction process
when the UE reaches the inner edge of cell (EoC) boundary of the serving DBS. The latter can be defined based on a distance threshold or a SS/SQ threshold. When the prediction is triggered, the SS and SQ prediction unit uses the stored measurements to predict SS and/or SQ of the serving DBS and the candidate DBSs.

The prediction analysis unit evaluates the predicted SS/SQ to determine if a certain DBS HO criteria is satisfied. If the predicted SS and/or SQ of a neighbouring DBS meets the HO criteria, then the prediction analysis unit queries the history prediction unit. The latter provides the prediction analysis unit with the probability of successful HO based on either the UE HO history or the DBS aggregated HO history. Based on these metrics as well as the predicted HO time, the prediction analysis unit may command the reporting unit to generate a new light-weight report called predictive measurement report (PrMR) and sends it to the serving DBS. This PrMR is sent only once as opposed to the periodic MR transmission in the conventional HO approach. At the serving DBS side, the HO mode decision unit evaluates the PrMR and commands the UE to operate either in a predictive mode or revert back to the conventional non-predictive mode. In the former, the conventional MR is suspended, the HO-related RAN signalling is performed in advance and HO control is delegated to the UE. On the other hand, the non-predictive mode follows the conventional HO procedure where the HO decision is taken by the serving DBS after the HO criteria is met. The operation, algorithms and interactions of these units are formulated and described in the following sections.

### 4.4.2 Signal Strength and Signal Quality Prediction Unit

Fig. 4.5 shows an exemplary operation of the SS and SQ prediction unit. It contains a short-term memory that stores the most recent $n$ active state measurements of the DBSs that reside within the angular span of the UE direction/speed. In other words, the SS/SQ prediction window size is $n$ measurements per DBS. The SS/SQ prediction is based on measurement trends. To minimise the UE storage requirements and remove signal fading/fluctuation effects, Grey system theory [137] is adopted as the trending approach. The Grey theory has been used in several fields, e.g., for disaster, season and

Figure 4.5: Exemplary operation of the signal strength and signal quality prediction unit

sequence prediction. It requires limited amount of input data and implicitly averages this data. The basic concept depends on translating the data sequence into a monotonic increasing function, representing this function by a differential equation and solving it to find the model’s parameters. For the problem under study (i.e., SS/SQ prediction), a GM(1,1)\(^3\) Grey model [137] can be constructed for each DBS as:

- The original SS/SQ measurements stored in the short-term memory are represented as a time series given by:

  \[
  y^{(0)}(i) = \left( y^{(0)}(1), y^{(0)}(2), ..., y^{(0)}(\bar{n}) \right), \quad i = 1, 2, 3, ..., \bar{n}, \tag{4.15}
  \]

  where the superscript \(0\) means original SS/SQ measurements (i.e., before processing) and \(i\) is the measurement index.

- An accumulated generating operation (AGO) translates \(y^{(0)}(i)\) to a monotonic\(^3\) the first 1 means the model uses first order differential equations, while the second 1 means there is one variable.
increasing function $y^{(1)}(i)$ as:

$$y^{(1)}(i) = \text{AGO} \left\{ y^{(0)}(i) \right\} = \left( y^{(0)}(1), \sum_{i=1}^{2} y^{(0)}(i), \sum_{i=1}^{3} y^{(0)}(i), \ldots, \sum_{i=1}^{n} y^{(0)}(i) \right).$$  \hspace{1cm} (4.16)

- Based on (4.16), an inverse accumulated generating operation (IAGO) can be formulated as:

$$y^{(1)}(i) = y^{(1)}(i - 1) + y^{(0)}(i).$$  \hspace{1cm} (4.17)

- The GM(1,1) model is defined by the following equation [137]:

$$\frac{dy^{(1)}}{du} + \ddot{a} y^{(1)} = \ddot{b},$$  \hspace{1cm} (4.18)

where $\ddot{a}$ is the develop parameter and $\ddot{b}$ is the grey input. The solution to (4.18) at time index $i$ is:

$$y^{(1)}(i + 1) = \left( y^{(1)}(1) - \frac{\ddot{b}}{\ddot{a}} \right) e^{-\ddot{a} i} + \frac{\ddot{b}}{\ddot{a}} = \left( y^{(0)}(1) - \frac{\ddot{b}}{\ddot{a}} \right) e^{-\ddot{a} i} + \frac{\ddot{b}}{\ddot{a}}. \hspace{1cm} (4.19)$$

By substituting the IAGO of (4.17) in (4.19), the predicted SS/SQ one measurement gap in advance $y^{(0)}_p(i + 1)$ can be expressed as:

$$y^{(0)}_p(i + 1) = e^{-\ddot{a} i} \left( 1 - e^{\ddot{a}} \right) \left( y^{(0)}(1) - \frac{\ddot{b}}{\ddot{a}} \right).$$  \hspace{1cm} (4.20)

Similarly, the predicted SS/SQ $j$ measurement gaps in advance can be formulated as:

$$y^{(0)}_p(i + j) = e^{-\ddot{a} (i+j-1)} \left( 1 - e^{\ddot{a}} \right) \left( y^{(0)}(1) - \frac{\ddot{b}}{\ddot{a}} \right).$$  \hspace{1cm} (4.21)

Equation (4.21) can be used to predicted a series of SS/SQ measurements. However, the model parameters $\ddot{a}$ and $\ddot{b}$ need to be calculated before the prediction is performed. These parameters can be obtained by expressing the derivative in (4.18) as:

$$\frac{dy^{(1)}}{du} \to y^{(1)}(i + 1) - y^{(1)}(i),$$  \hspace{1cm} (4.22)

and the right hand side of (4.22) can be replaced with $y^{(0)}(i + 1)$ based on the IAGO of (4.17), i.e.,

$$\frac{dy^{(1)}}{du} \to y^{(0)}(i + 1).$$  \hspace{1cm} (4.23)

The mean value of adjacent SS/SQ measurements is:

\[ z^{(1)}(i) = \frac{1}{2} y^{(1)}(i) + \frac{1}{2} y^{(1)}(i - 1) \rightarrow y^{(1)}(u). \tag{4.24} \]

Based on (4.23) and (4.24), the Grey differential equation of (4.18) can be rewritten as:

\[ y^{(0)}(i) + \bar{a} z^{(1)}(i) = \bar{b}. \tag{4.25} \]

Rearranging (4.25) and writing the resultant equation in a matrix form yields:

\[
\begin{bmatrix}
  y^{(0)}(2) \\
  y^{(0)}(3) \\
  \vdots \\
  y^{(0)}(\tilde{n})
\end{bmatrix}
= 
\begin{bmatrix}
  -z^{(1)}(2) & 1 \\
  -z^{(1)}(3) & 1 \\
  \vdots & \vdots \\
  -z^{(1)}(\tilde{n}) & 1
\end{bmatrix}
\cdot
\begin{bmatrix}
  \bar{a} \\
  \bar{b}
\end{bmatrix},
\tag{4.26}
\]

Finally, \( \bar{a} \) and \( \bar{b} \) can be obtained by solving (4.26), i.e.,

\[
\begin{bmatrix}
  \bar{a} \\
  \bar{b}
\end{bmatrix}
= 
\begin{bmatrix}
  -z^{(1)}(2) & 1 \\
  -z^{(1)}(3) & 1 \\
  \vdots & \vdots \\
  -z^{(1)}(\tilde{n}) & 1
\end{bmatrix}^{-1}
\cdot
\begin{bmatrix}
  y^{(0)}(2) \\
  y^{(0)}(3) \\
  \vdots \\
  y^{(0)}(\tilde{n})
\end{bmatrix},
\tag{4.27}
\]

The SS and SQ prediction unit uses this model to predict SS and SQ of the serving DBS and a subset of the top-\( \bar{n} \) other detectable DBSs. These results are fed to the prediction analysis unit. It is worth mentioning that other trending techniques, such as polynomial fitting or sample extrapolation, can be used instead of the Grey model. For instance, the fitting technique proposed in Section 3.2.1 and formulated by (3.6) and (3.7) to predict the zero lag CC based on noisy channel estimate, can be reused here as the trending approach.

The expected HO time can be predicted based on rate of SS/SQ degradation. The measurement prediction is a time-series prediction, and it is performed on a sample-basis. Thus if the measurement gap \( \delta_g \) is constant (such as in current standards), the SS and SQ prediction unit predicts a series of measurements until the HO criteria is satisfied (see Fig. 4.5). For example, if the prediction is performed for \( I_p \) samples in advance (i.e., assuming that the current measurement index is \( \bar{n} \), and the predicted

\[ d_{s,thr} \]

\[ y_{s,thr} \]

\[ 0 \]

Serving DBS

CoC

EoC

Inner boundary

Outer boundary

Figure 4.6: Methods for UE location estimation and CoC/EoC prediction triggering threshold

SS/SQ that satisfies the HO criteria has index \( \hat{n} + I_p \), then the predicted remaining time for HO is:

\[
\text{Predicted HO time} = I_p \cdot \delta_g. \tag{4.28}
\]

4.4.3 Location and Span Estimation Unit

This main objective of this unit is minimising the UE processing load and storage requirements. The prediction process in Section 4.4.2 can be continuously executed until a target DBS is found. However, such a continuous operation may not be feasible from battery and processing perspectives. A more convenient design approach is to trigger the prediction process when a certain triggering criteria is satisfied. This suggests a two boundary DBS cell structure, where the prediction is triggered at the inner boundary while the actual HO is performed at the outer boundary. Fig. 4.6 shows two approaches that can be used by the location and span estimation unit to trigger the prediction process at the inner boundary, based on the UE location w.r.t. the serving DBS, i.e., centre of cell (CoC) or EoC. Notice that the CoC/EoC classification is based on the inner boundary.

These approaches include position-based distance calculation and serving DBS signal measurements. The former requires the serving DBS to broadcast its location, e.g., in the form of longitude and latitude. Then, the distance between the UE and the serving DBS can be calculated based on the UE position (provided by either a GPS or other positioning techniques). When this distance equals to or greater than a certain
An appropriate setting of $d_{s,thr}$ and/or $y^{(0)}_{s,thr}$ is of great importance in improving the performance of the proposed scheme. A large $d_{s,thr}$ (i.e., low $y^{(0)}_{s,thr}$) setting may result in a too late prediction, i.e., the HO may happen before the prediction process is triggered. On the other hand, a small $d_{s,thr}$ (i.e., high $y^{(0)}_{s,thr}$) setting may lead to a too early prediction. This in turn increases the error probability, due to the large gap between the time when the prediction is performed and the time when the actual HO happens. In addition, radio channel and/or UE direction will have a higher changing probability when the actual HO happens. As an illustrative example, Fig. 4.7 shows simulation results for the measurement prediction precision with several prediction advance periods. As can be noticed, the prediction of the $i^{th}$ SS/SQ measurement is more accurate than the prediction of the $j^{th}$ SS/SQ measurement, where $i < j$.

Assuming a constant speed $V$ and a hysteresis-based HO criteria, i.e., the HO happens if the following condition is true:

$$\log \left( y^{(0)}_n \right) \geq \log \left( y^{(0)}_s \right) + \log (\Theta),$$

(4.29)

where \( y_s^{(0)} \) and \( y_n^{(0)} \) are the SS/SQ of the serving and the neighbouring DBSs, respectively, \( \Theta \) is the HO hysteresis, and the parameters in (4.29) are in linear scale. For a general path loss model \( \chi R^\xi \), where \( \chi \) is the distance-independent path loss component, \( R \) is the distance between the Tx and the Rx, and \( \xi \) is the path loss exponent. Then it can be proved that the actual HO happens after \( I_p \) measurements referenced to the prediction triggering point, where \( I_p \) is expressed as:

\[
I_p = \left( \frac{\frac{q_s \Theta}{q_n} \frac{\psi}{\cos(\phi_n)}}{V \delta_g \left( 1 + \left( \frac{\cos(\phi_s)}{\cos(\phi_n)} \right) \left( \frac{q_s \Theta}{q_n} \right)^\frac{\xi}{\xi} \right)} \right) - d_{s.thr},
\]

(4.30)

where \( q_s \) and \( q_n \) are the transmit power of the serving and the neighbouring DBSs, respectively, \( \psi \) is the inter-site distance, \( \phi_s \) is the angle between the line connecting the DBSs and the line connecting the serving DBS with the UE location when the HO happens, \( \phi_n \) is the angle between the line connecting the DBSs and the line connecting the neighbouring DBS with the UE location when the HO happens. It can be noticed that (4.30) depends on the UE speed. Thus using a DBS-based unified triggering threshold for all users implies that different users will have different \( I_p \) values. Expressed differently, if the prediction is triggered at the same location for all users, then low speed users will have to predict more measurements than high speed users. This may increase the error probability and decrease the prediction precision as shown in Fig. 4.7. This suggests a UE-specific prediction triggering threshold \( d_{s.thr,UE} \) that takes into account network parameters as well as UE parameters. This threshold can be obtained by solving (4.30) for \( d_{s.thr} \), i.e.,

\[
d_{s.thr,UE} = \left( \frac{q_s \Theta}{q_n} \right)^\frac{\xi}{\xi} \left( \frac{\psi - d_{s.thr} \cos(\phi_s)}{\cos(\phi_n)} \right) - I_p V \delta_g \left( 1 + \left( \frac{\cos(\phi_s)}{\cos(\phi_n)} \right) \left( \frac{q_s \Theta}{q_n} \right)^\frac{\xi}{\xi} \right),
\]

(4.31)

The UE angular span is utilised to narrow down the candidate DBS set. A high speed user usually has a smaller span (i.e., probability of changing the direction is small) as compared with a low speed user. This allows the former to store and process measurements of a smaller number of DBSs as compared with the latter. The location and span estimation unit calculates the UE angular span \( \Omega \) by

\[
\Omega = 2 \pi e^{-\eta V},
\]

(4.32)
where $\eta$ is the span gradient. Different DBSs can define different values for $\eta$ which can be learned from the users’ behaviour. For instance, a highway DBS may define a large gradient which results into a small span (i.e., a low mobility in a highway may be attributed to traffic conditions rather than to a direction change intention). A new span is defined when the UE changes its speed abruptly, when the UE changes its direction by an angle larger than the initial span, or at regular time intervals. Based on the location of the top-$\bar{n}$ detectable DBSs, the location and span estimation unit selects the DBSs that reside within the UE span as candidates for the prediction process. This result is fed to the SS and SQ prediction unit.

### 4.4.4 History Prediction Unit

The history prediction unit provides statistical historical information that helps the prediction analysis unit to confirm or reject the measurement-based HO prediction. It calculates the HO probability from the serving DBS to the predicted DBS based on either the aggregated NL HO history [104], [138] or the UE HO history. The former is already available in current standards at the network side in the form of a HO frequency table, and it provides statistical information based on the crowd behaviour. Table 4.1 provides an illustrative example of the NL HO history in a table format, where $C_{i,j}$ is the NL-based aggregated HO count from DBS$_i$ to DBS$_j$. Typically, each row in Table 4.1 is maintained by the source DBS. This NL HO history can be translated into a transition probability. For instance, the NL-based transition probability $t_{i,j}$ from

DBS_i to DBS_j can be obtained by:

\[ t_{i,j} = \sum_{u, v \in N_i} \frac{C_{i,j}}{C_{i,u}}, \]  \hspace{1cm} (4.33)

where \( N_i \) has the same definition as in Section 4.3.1.

The NL-based history can be used for DBSs covering areas characterised by high batch HO rates. For example, where multiple UE on a train perform simultaneous HOs. However, the NL-based approach may not be suitable for individual users since it provides a coarse and less accurate estimate based on the crowd behaviour. This suggests a UE-based approach where individual users maintain separate statistical information based on their own history. The history-based prediction scheme provided in Section 4.3 can be used for this purpose. In other words, the history prediction unit in Fig. 4.4 follows the learning and prediction scheme of Section 4.3.

4.4.5 Prediction Analysis and Handover Mode Decision Units

The prediction analysis unit evaluates the predicted SS/SQ to determine if a certain HO criteria is satisfied. The latter is left as an implementation aspect to ensure a generic prediction scheme that does not depend on a particular HO model. For instance, the condition of (4.29) can be used as an example for the HO criteria. A HO hysteresis needs to be applied to the trend to ensure that prevailing conditions only are acted on to avoid HO ping-pong. The prediction analysis unit confirms/rejects the measurement-based prediction based on the UE HO history and the predicted HO time. Consider \( t_{s,p} \) as the transition probability from the serving DBS_s to the measurement-based predicted DBS_p, based on either the crowd behaviour or the individual user behaviour as explained in Section 4.4.4. Define \( t_{min} \) as the minimum HO probability to confirm the prediction from a history perspective. Then the prediction analysis unit confirms the measurement-based prediction if the following condition is true:

\[ t_{s,p} \geq t_{min}, \]  \hspace{1cm} (4.34)

and it commands the reporting unit to send a PrMR, which contains the predicted DBS along with the predicted remaining time for HO given by (4.28). Otherwise, the

The HO mode decision unit decides the type of HO to be followed by the UE, i.e., predictive or non-predictive HO. It can be located at the UE side and integrated with the prediction analysis unit, thus the outcome of the latter implicitly defines the HO type. If the final decision has to be taken based on additional policies defined by the network (i.e., network-controlled UE-assisted decision), then the HO mode decision unit can be moved to the DBS side as shown in Fig. 4.4.

It is worth mentioning that the proposed predictive schemes have different use cases.
and they depend on different parameters that can not compared with each other. The history-based predictive scheme depends on the UE HO history to predict future HO events. Irrespective of the instantaneous network conditions, interference levels, UE direction and MRs, the history-based predictive scheme produces the same outcome. For instance, even if the UE changed its direction (which implies that the received DBS power, interference and MRs will change), the predicted DBS in the history-based predictive scheme will remain the same. On the other hand, the measurement-based context-aided predictive scheme depends on instantaneous parameters such as MRs (which take into account the received DBS power and interference), the UE direction, the UE speed and the HO hysteresis. Consequently, the performance of this scheme is directly linked to the instantaneous network conditions, HO parameters and UE context.

4.5 Performance Evaluation

4.5.1 History-based Predictive Handover Results

HO traces have been generated to assess performance of the history-based predictive DBS HO scheme. The considered network topology consists of 69 hexagonal shaped DBSs. Traces for 100 consecutive days are collected where the trajectory of each day consists of 10 HOs. Several mobility scenarios are considers: 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.

Prediction accuracy

The evaluation is based on the prediction accuracy which is defined as the ratio between the number of correctly predicted DBSs and the total number of the DBSs visited by
4.5. Performance Evaluation

Fig. 4.9 provides the average prediction accuracy for different values of $\epsilon$. In the regular movement scenario (i.e., 0% random mobility), the proposed scheme predicts the trajectory with an accuracy of 97 – 99% for $\epsilon \geq 0.3$. However, the prediction accuracy decreases at smaller values of $\epsilon$ to reach 92% with $\epsilon = 0.1$. This can be traced to the fact that small $\epsilon$ values require more observations (i.e., a longer history) before a reliable decision can be made. It can be noticed that the prediction accuracy is roughly 0% when $\epsilon = 0$. In the latter case, each trajectory does not have any effect on the updated $T$ as depicted by (4.9) and (4.10). Expressed differently, the model does not build any history for the user, hence the prediction is independent of the mobility history. It is worth mentioning that the evaluation is performed without a training data set. Since the prediction is dependent on the historical trajectory, the accuracy in the 1st day is roughly 0% irrespective of the $\epsilon$ value. As a result, the average accuracy in Fig. 4.9 does not reach 100%.

With 10% random mobility, the average prediction accuracy is $\approx 86\%$ for $\epsilon$ in the range $[0.2, 0.7]$. However, the accuracy decreases to $\approx 80\%$ when $\epsilon > 0.7$. Since the user performs some random movements, a high dependency on the most recent trajectory slightly decreases the prediction accuracy. A similar behaviour can be observed in the 20% and the 30% random mobility scenarios where the prediction accuracy reaches a
peak of 77\% and 68\%, respectively, when \( \epsilon \) in the range [0.2, 0.4] and decreases at larger values of \( \epsilon \). In other words, reducing \( \epsilon \) minimises the effect of random movements on the updated \( T \), but it increases the required observation period. Thus it can be said that \( \epsilon \) is an important UE-specific parameter that can be tuned for each user to reach a particular accuracy target.

It is worth mentioning that the predictive HO schemes proposed in this chapter aim to reduce the HO latency rather than reducing the HO signalling overhead. This is achieved by exploiting the prediction outcome to perform the HO preparation phase in advance before the HO criteria is met. The gains of this approach in terms of reduction in HO latency are shown and discussed at the end of this section. However, the HO signalling overhead (in terms of the number of the HO-related RRC messages) with correct prediction remains the same as the overhead without HO prediction. In the case of incorrect prediction, the HO signalling overhead becomes higher than the overhead without HO prediction. This can be traced to the signalling model where additional RRC messages are needed to cancel the reserved resources at the incorrectly predicted DBS, as can be seen in Fig. 4.1. To quantify this penalty, Table 4.2 shows the normalised predictive HO signalling overhead (in terms of the number of the HO-related RRC messages normalised with the number of the HO-related RRC messages in the conventional non-predictive HO scheme) vs the prediction accuracy.

**Entropy-based handover mode selection statistics**

Fig. 4.10 provides the cumulative distribution function (CDF) of the prediction entropy for both single and multiple HOs predictions. It can be noticed that the entropy of a single HO prediction is significantly less than the entropy of multiple HOs 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\textsuperscript{th} percentile of 0.22 hartley and 0.96 hartley for single and multiple HOs 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. The random mobility
4.5. Performance Evaluation

Table 4.2: Normalised predictive handover signalling overhead vs prediction accuracy

<table>
<thead>
<tr>
<th>Prediction accuracy</th>
<th>Normalised overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>1.18</td>
</tr>
<tr>
<td>10%</td>
<td>1.16</td>
</tr>
<tr>
<td>20%</td>
<td>1.15</td>
</tr>
<tr>
<td>30%</td>
<td>1.13</td>
</tr>
<tr>
<td>40%</td>
<td>1.11</td>
</tr>
<tr>
<td>50%</td>
<td>1.09</td>
</tr>
<tr>
<td>60%</td>
<td>1.07</td>
</tr>
<tr>
<td>70%</td>
<td>1.05</td>
</tr>
<tr>
<td>80%</td>
<td>1.04</td>
</tr>
<tr>
<td>90%</td>
<td>1.02</td>
</tr>
<tr>
<td>100%</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 4.10: CDF of the history-based prediction entropy. Acronym RM: Random Mobility

effect can also be seen in Fig. 4.10, where the 90\textsuperscript{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. A similar behaviour can also be seen in the multiple HOs prediction case. Expressed differently, the prediction confidence decreases as the UE randomness increases.
4.5. Performance Evaluation

Figure 4.11: Statistics of the actual executed handover type vs $h_{thr}$, with 30% random mobility

The effect of $h_{thr}$ on the switching point between predictive and non-predictive HOs can be seen in Fig. 4.11 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. 4.11a that 50% of the predictions satisfy the predictive HO condition (i.e., $h(p_1) \leq h_{thr}$). This can be linked to the high confidence (i.e., low entropy) of the single HO prediction depicted by Fig. 4.10. 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. 4.11b shows that all of the multiple HOs predictions do not satisfy the predictive HO condition when $h_{thr} \leq 0.5$ hartley due to the high entropy of this case. It can be observed that as $h_{thr}$ increases (i.e., the confidence requirement decreases), more predictions satisfy the advance DBS signalling criteria at the cost of a slight increase in the penalty (i.e., the incorrect predictions that follow the predictive HO signalling procedure).

Although the single HO prediction provides a higher confidence and a better performance than the multiple HOs prediction, the latter is more beneficial in energy-efficient CDSA networks with DBS sleep modes. Predicting several HOs in advance allows optimised DBS on/off decisions rather than instantaneous decisions with on/off oscillations. As an illustrative example, consider a UE with a predicted DBS HO sequence given by: $\text{DBS}_1 \rightarrow \text{DBS}_2 \rightarrow \text{DBS}_3 \rightarrow \text{DBS}_4$. When $\text{DBS}_3$ is not serving any UE, it can
4.5. Performance Evaluation

Table 4.3: Latency values for handover signalling messages

<table>
<thead>
<tr>
<th>Latency description</th>
<th>Value (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmission latency between DBSs over X2</td>
<td>5</td>
</tr>
<tr>
<td>Transmission latency between UE and DBS</td>
<td>6.5*</td>
</tr>
<tr>
<td>Transmission latency between DBS and MME</td>
<td>8.5</td>
</tr>
<tr>
<td>Processing latency at DBS</td>
<td>4</td>
</tr>
<tr>
<td>Processing latency at MME</td>
<td>5**</td>
</tr>
<tr>
<td>Processing latency at S-GW</td>
<td>5**</td>
</tr>
<tr>
<td>Latency to detach from the source DBS and access the target DBS</td>
<td>12</td>
</tr>
</tbody>
</table>

* Includes processing.

** Does not include UE context retrieval of 10 ms.

be switched off for energy saving. However, DBS$_3$ has to be switched on before the UE reaches this DBS. Depending on the DBS activation and deactivation delays, it might be more appropriate to keep DBS$_3$ on to avoid the on/off oscillations. This can be achieved by including the predicted DBS HO sequence in the decision process. Expressed differently, the multiple HOs prediction enables a better planning for DBS on/off decisions.

**Handover latency**

In the following, potential benefits of the history-based predictive DBS HO scheme are evaluated in terms of signalling latency. For simplicity, the approach of [133] has been followed by assuming that the transmission delay for different messages between the same source-destination pair is the same irrespective of the message size. Similarly, the processing delay for different messages at the same node is constant. In addition, it has been assumed that the MME and the S-GW are located in the same location, thus the transmission delay between these nodes is negligible. Table 4.3 provides the latency values which are based on the feasibility study reported in [131] for the intra-LTE X2 HO procedure.

Fig. 4.12 shows the HO signalling latency as a function of the entropy-based switch-
4.5. Performance Evaluation

Figure 4.12: Signalling latency vs $h_{thr}$ in conventional handover and history-based predictive DBS handover. Acronym RM: Random Mobility

With a conservative confidence threshold of $h_{thr} \leq 0.08$ hartley, the proposed scheme generates the same signalling as the conventional HO scheme. This can be linked to Fig. 4.11 where the high confidence requirement rejects all the predictions, thus all the HOs follow the non-predictive procedure. Increasing $h_{thr}$ enables more predictions to satisfy the triggering condition, which in turns reduces the signalling latency of the history-based predictive DBS HO scheme by up to 34% for the regular mobility scenario with $h_{thr} \geq 0.3$ hartley, and by 23% for the 30% random mobility scenario with $h_{thr} \geq 0.6$ hartley. Nonetheless, it is worth emphasising that increasing $h_{thr}$ increases the risk of incorrect predictions. Since the UE randomness and the prediction entropy are proportional to each other as depicted by Fig. 4.10, users with regular mobility profiles can achieve the maximum gains with a lower risk (i.e., lower $h_{thr}$ setting) as compared with random mobility users. 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 history-based predictive DBS HO scheme.

4.5.2 Measurement-based Context-aided Predictive Handover Results

Prediction triggering threshold analysis

Fig. 4.13 shows effect of the UE speed on the expected HO time (in measurement gaps) referenced to the prediction triggering point, with a unified triggering threshold for
4.5. Performance Evaluation

The UE needs to correctly predict the measurement reports up to \( j \) gaps away. Hysteresis = 2 dB, measurement gap = 200 ms, ISD = 130 m, serving power = 38 dBm, neighbour power = 38 dBm.

Hysteresis = 0 dB
Hysteresis = 1 dB
Hysteresis = 2 dB
Hysteresis = 3 dB
Hysteresis = 4 dB

Fig. 4.13b indicates that the speed effect on the expected HO time (with a unified triggering threshold) is significantly influenced by the HO hysteresis. It can be seen that increasing the hysteresis \( \Theta \) increases the slope (in absolute value) of the HO time vs speed graph when \( d_{s,thr} \) is constant for all users. This indicates that a low hysteresis setting may be appropriate when \( d_{s,thr} \) is unified for all users. Nonetheless, the HO hysteresis provides other benefits such as delaying the HO to avoid HO ping-pongs and
it removes the SS/SQ fluctuation effects. As a result, decreasing Θ may come at the expense of increasing HO ping-pongs rates.

The observations in Fig. 4.13 motivate a UE-specific prediction triggering threshold, which is provided in Fig. 4.14. It can be noticed that $d_{s.thr.UE}$ is inversely proportional to the UE speed. Expressed differently, low speed users start the prediction process at a larger distance from the serving DBS as compared with high speed users. This ensures that all users predict the same number of SS/SQ measurements, which in turn allows to set a maximum advance period in order to control the prediction precision and error probability. For example, to ensure that the prediction process does not start more than 8 measurement gaps before the actual HO happens, then a user with $V = 5$ km/hr and $V = 80$ km/hr triggers the prediction process at 70 m and 37 m, respectively, from the serving DBS as shown in 4.14a. Since the expected HO time is inversely proportional to the distance from the serving DBS, then increasing the advance period $I_p$ reduces $d_{s.thr.UE}$. On the other hand, Fig. 4.14b indicates that the UE-specific $d_{s.thr.UE}$ and the HO hysteresis have a proportional relationship. This can be linked to the fact that the hysteresis delays the actual HO. Hence, for a fixed advance prediction period target, a higher hysteresis setting increases the required prediction triggering distance from the serving DBS.
Table 4.4: Simulation parameters of the measurement-based context-aided predictive DBS handover scheme

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network layout</td>
<td>Hexagonal grid, 19 omnidirectional DBSs</td>
</tr>
<tr>
<td>Inter-site distance</td>
<td>130 m</td>
</tr>
<tr>
<td>DBS transmit power</td>
<td>38 dBm</td>
</tr>
<tr>
<td>Transmit mode</td>
<td>SISO (Single Input Single Output)</td>
</tr>
<tr>
<td>User density</td>
<td>5 UE/DBS</td>
</tr>
<tr>
<td>User speed</td>
<td>10 km/hr for 100% of the users</td>
</tr>
<tr>
<td>Channel model</td>
<td>3GPP Typical Urban [140]</td>
</tr>
<tr>
<td>Path loss model</td>
<td>3GPP Urban [141]</td>
</tr>
<tr>
<td>Frequency</td>
<td>2 GHz</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Scheduler</td>
<td>Round robin</td>
</tr>
<tr>
<td>Measurement gap</td>
<td>200 ms</td>
</tr>
</tbody>
</table>

Prediction statistics and gains

A second set of system level simulations has been performed to assess performance and gains of the measurement-based context-aided predictive DBS HO scheme. Table 4.4 provides the considered simulation parameters which are mostly aligned with the assumptions in [139]. Fig. 4.15 shows prediction accuracy and statistics for several SS triggering thresholds and HO hysteresis values. It can be noticed that this scheme provides a 90% prediction accuracy when $y_{s,thr}^{(0)} \geq -62$ dBm and HO hysteresis is used (i.e., $\Theta > 0$ dB). In addition, it significantly reduces the percentage of incorrect predictions that are not rejected by the prediction analysis unit. Precisely, only $2.5\% - 9.6\%$ of the predictions resulted in HOs to DBSs other than the actual target DBSs. A very low SS triggering threshold of $y_{s,thr}^{(0)} = -64$ dBm with a low (or no) HO hysteresis setting results in a significant number of too late predictions. This can be traced to the fact that a low (or no) hysteresis results in an early HO while a low SS triggering threshold
4.6 Conclusion

Predictive HO signalling at DBS-level has been proposed in this chapter. With the main objective of minimising the CDSA HO-related RAN signalling and the associated latency, two predictive schemes have been proposed. These include a history-based predictive DBS HO and a measurement-based context-aided predictive DBS HO. Both schemes include proactive HO mode selection models to minimise the HO signalling
4.6. Conclusion

Figure 4.16: Handover signalling latency reduction in the measurement-based context-aided predictive DBS handover scheme w.r.t. the conventional handover latency, since the predictive HO management strategies may not be suitable in some cases, e.g., unpredictable users with highly random mobility profiles or users visiting new DBSs. Considering the dual connectivity feature of the CDSA, such predictive approaches can be applied with relaxed constraints.

The history-based predictive DBS HO model relies on past HO history to predict future HO events. It adopts the DTMC to model this history in a probabilistic manner. An online learning process is proposed to update the DTMC transition matrix without maintaining HO frequency tables, resulting in a memory/storage friendly prediction scheme. In addition, a recent trajectory dependency parameters $\epsilon$ is proposed to control the model’s reaction to random and less frequent movements. It has been found that this parameter can significantly reduce the effect of random movements and it improves the prediction performance, where the peak accuracy can be achieved through an inversely proportional relationship between $\epsilon$ and the UE mobility randomness. The switching point between predictive and conventional HO modes is defined based on the prediction confidence, which is measured by the entropy. The latter is proportional to the UE mobility randomness as depicted by simulations and it controls the gain/risk ratio. Since the history-based predictive scheme does not depend on UE-DBS measurements, it can also be used in solving the DP hole problem in green SON-based CDSA
implementation with dynamic on/off DBS operation.

The measurement-based context-aided predictive DBS HO scheme is operated at the UE, and it provides additional information such as the expected HO time in addition to the target DBS. It combines physical proximity (i.e., location information at the UE) to a virtualised UE view of DBS coverage, RF performance derived from SS/SQ measurements, context information and HO history. The SS/SQ measurements are modelled as a time series in a Grey fashion to predict future HO events and the remaining time for HO. In addition, the UE speed and direction are utilised to minimise the storage and processing requirements by narrowing down the candidate DBS set based on the UE angular span. A UE-specific prediction triggering threshold is formulated to improve the measurement prediction precision whilst minimising the UE processing load. For a certain advance prediction period (and hence a certain precision target), it has been found that the UE-specific triggering threshold is inversely proportional (in distance format) and directly proportional (in SS/SQ format) to the UE speed. The switching point between predictive and conventional HO procedures is defined based on the successful HO probability which is obtained from the history-based prediction model.

The LTE X2 HO procedure is considered as a benchmark for the conventional non-predictive HO strategy. In addition, the standard procedure is adapted to the predictive scenario in order to evaluate its signalling latency. The latter is modelled as a function of the transmission and the processing delays for each signalling message involved in the HO process. Simulation results show that the proposed predictive schemes reduce the HO signalling latency by 34% as compared with the conventional LTE X2 HO procedure. In the history-based predictive scheme, users with regular mobility profiles can achieve the highest gains with a lower risk while random mobility users need to assume a higher risk to increase the potential gains. On the other hand, the highest gains in the measurement-based context-aided predictive scheme depends on network-defined HO parameters (such as the HO hysteresis), in addition to the UE-specific prediction parameters (such as the prediction triggering threshold). Although the peak gains of both schemes are roughly the same, each one has its own inputs, granularity and use cases.
Chapter 5

Core-network Transparent
Handover Signalling and
Backhaul Latency Constraints

This chapter focuses on the HO-related CN signalling and proposes a CN-transparent signalling scheme for intra-CBS HOs with minimal CN signalling overhead. An analytical framework is developed to model the CN signalling load of the proposed scheme as a function of network density, user mobility and session characteristics. This scheme is constrained by the DP-BL. Thus, an analytical model is derived to evaluate the impact of several parameters that can be tuned to reach a particular latency target. Furthermore, closed-form expressions are obtained for the densification limits under latency constraints. Moreover, the CN-transparent signalling model is integrated with the predictive HO signalling techniques of Chapter 4 to evaluate the overall gains of the integrated scheme. The work presented in this chapter has been published in [11], [14] and [15].

5.1 Introduction

The CDSA system model, where active UE maintain a dual connection with both the CBS and the DBS, enables adaptive and efficient signalling mechanisms as discussed
in Chapters 2–4. From a mobility signalling perspective, the predictive models in Chapter 4 show promising gains over the conventional approach. These predictive HO techniques enable advance DBS signalling, which in turn reduces the HO latency and the associated RAN and air interface overhead. However, the CN signalling remains the same with/without mobility prediction. This can be traced to the fact that the HO procedure (with/without mobility prediction) generates signalling towards the CN to transfer all channels (i.e., control and data) from one BS to another. In the conventional RAN architecture, this CN signalling can not be avoided since the active UE maintain a single connection with the serving BS only, resulting in a CN-visible HO. The CDSA, however, could offer a solution to this problem by exploiting the dual connectivity feature and redesigning the DP routing paths.

The authors of [16], [95] and [96] argue that the CDSA does not require changing the CP link as long as the UE mobility is within the same CBS. This in turn could result into improving the mobility KPIs. A similar approach has been adopted in [5] and [6] where several mobility enhancement opportunities are discussed. The models in [142] depend on simulations to analyse HOF rate of the CDSA by focusing on the fixed CP link feature. It is worth mentioning that most of the work in this area provide a qualitative discussion rather than a proper analysis with quantitative results. In addition, the state-of-the-art proposals focus on a particular aspect of dual connectivity: mobility performance is reflected by the CBS-level mobility because the UE is always anchored to the CBS. Although this feature improves the HO KPIs such as failure rates, it does not reduce the signalling overhead. The latter is linked to the DP routing path, which itself depends on the adopted separation 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. Expressed differently, the intra-CBS HOs require changing the UE-DBS link without changing the UE-CBS link. On the other hand, the inter-CBS HO requires changing both the UE-DBS link and the UE-CBS link, since it is performed between DBSs with different CBS anchor points. As discussed in Section 4.1, the HO-related CN signalling is mainly used to switch the DP path when the user performs a HO in the conventional architecture. This chapter progresses beyond the state-of-the-art by exploiting the
CBS as a mobility anchor point for the UE and as a DP anchor point for the DBSs. Therefore, the CN-CBS DP path remains the same as long as the user mobility is within the same CBS. Although the intra-CBS HOs require changing the DBS, the CBS-DBS DP path can be switched locally at the CBS. Thus the intra-CBS HOs do not generate CN signalling, resulting in a CN-transparent HO signalling scheme. Notice that the CDSA and the dual connectivity concepts are still maintained because the UE receives data transmission and control signalling form the DBS and from the CBS, respectively.

This CN-transparent HO scheme can significantly reduce the HO-related CN signalling load as compared with the conventional CN-visible HO model. To evaluate these potential benefits, an analytical framework is developed to model the HO-related CN signalling load in both schemes, i.e., the proposed CN-transparent HO and the conventional CN-visible HO. 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 CN signalling under general distributions in both schemes. In addition, closed-form expressions are obtained in a special case where session duration and cell residence time are exponentially distributed. Unlike the predictive HO models that aim to reduce the HO latency, the CN-transparent HO model focuses on reducing the HO signalling overhead. This model provides a comparison criteria that can be used to assess the claimed gains in terms of saving in CN signalling overhead.

Despite the potential gains, the CN-transparent HO signalling with DP routing through both the CBS and the DBS induces an additional DP latency as compared with the conventional CN-visible HO. To investigate this penalty, a stochastic geometry model is proposed to derive closed-form expressions for the DP-BL under both HO schemes. This model is used to analyse the impact of several parameters, such as DBS and CBS densities, data load and processing capabilities, on the DP-BL. In addition, an upper bound for the DBS density under DP-BL constraints is derived. This upper bound is used to assess the CN-transparent HO gains in dense deployment scenarios without violating the DP-BL constraints.

The reminder of this chapter is structured as follows: Section 5.2 introduces a high
level overview of the proposed system model and describes the stochastic geometry modelling approach. Section 5.3 develops the CN-transparent HO signalling scheme, and derives closed-form expressions for the CN signalling probability and load. Section 5.4 models the DP-BL and derives the upper bound of the DBS density. Section 5.5 presents and discusses numerical and simulation results that assess performance of the proposed scheme, while Section 5.6 concludes the chapter.

5.2 System Model

The proposed CN-transparent HO signalling scheme utilises the large footprint of the CBS to minimise the CN signalling load. In this model, the CBS is used as a DP anchor point for the DBSs. Subsequently, 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 alleviates the CN signalling since the path switching is performed locally at the CBS. The reduction in the CN signalling load minimises the HO overhead and contributes towards reducing the HO latency. As discussed and modelled in Chapter 4, this latency is highly dependent on the signalling procedure since each HO signalling message requires some time to be prepared, transmitted and processed at the destination. In order to evaluate the overall gains of both schemes, the CN-transparent HO model is integrated with the predictive HO schemes of Chapter 4. Fig. 5.1 shows a high level overview of the integrated HO signalling model.

This chapter focuses on modelling the CN-transparent HO part of the integrated model, since the predictive part has been modelled and analysed in Chapter 4. To model the CN-transparent HO signalling scheme analytically, stochastic geometry and Poisson Point Process (PPP) are adopted as the modelling approach. The tractability, lack of edge effects and natural inclusion of different cell sizes have made the PPP a popular modelling approach in cellular research community. It has been adopted in [143] to derive coverage probability and DL throughput of homogeneous cellular networks, and in [144] to model the inter-cell signalling overhead in HetNets. The authors of [145] use the PPP model to analyse the impact of user density on the SINR distribution. Similarly, [146] optimises the UE-BS association in HetNets by using the PPP to model
5.2. System Model

*Integrated Handover Signalling Model*

- **Scheme**
  - CN-Transparent Handover Signalling
  - Predictive Handover Signalling

- **Enabler**
  - CBS as Mobility and DP Anchor Point
  - DBS with Advance RAN Signalling

- **Benefit**
  - Minimal CN Handover Signalling Load
  - Minimal Handover Signalling Latency

Figure 5.1: High level overview of the integrated predictive and core-network transparent handover signalling scheme

Both user and network nodes. Following these mature studies, the proposed framework adopts the PPP to model the CDSA nodes.

The system model consists of DBSs, CBSs and CN aggregation nodes (ANs) that are modelled as three independent PPPs in $\mathbb{R}^2$: $\Phi_d$ with density $\lambda_d$ for the DBSs, $\Phi_c$ with density $\lambda_c$ for the CBSs and $\Phi_a$ with density $\lambda_a$ for the ANs, where $\lambda_d > \lambda_c > \lambda_a$. $d$, $c$ and $a$ are arbitrary points of $\Phi_d$, $\Phi_c$ and $\Phi_a$ respectively. This model results in voronoi spanning trees, where any realisation of $\Phi_d$, $\Phi_c$ and $\Phi_a$ yields random trees as shown conceptually in Fig. 5.2. The CDSA network configuration and DP backhaul path depend on the HO model, which in turn defines the PPP connections. For each HO scheme, the DP path and the PPP connections can be identified as:

- **CN-transparent handover signalling**: the RRC connection is maintained at the CBS where the CP/DP separation is performed. Thus the data path switching is performed locally at the CBS for intra-CBS HOIs rather than triggering a switch request towards the CN. Each DBS is connected to a single CBS, where the CBS-DBS association is based on least distance, i.e., $\text{DBS}_i$ is connected to $\text{CBS}_j$ if and only if $\text{DBS}_i$ is in the voronoi cell of $\text{CBS}_j$. In addition, each CBS is connected to a single AN based on the least distance rule, i.e., $\text{CBS}_j$ is connected to $\text{AN}_k$ if and only if $\text{CBS}_j$ is in the voronoi cell of $\text{AN}_k$. 
• **CN-visible handover signalling:** each DBS acts as a DP anchor point and has a separate RRC entity. Consequently, the CP/DP separation is performed at the CN. Thus all HOs (i.e., intra- and inter-CBS HOs) are visible to the CN since they require data path switching operation at the CN. Each DBS is connected directly to a single AN, where DBS<sub>i</sub> is connected to AN<sub>k</sub> if and only if DBS<sub>i</sub> is in the voronoi cell of AN<sub>k</sub>. In addition, each CBS is connected to a single AN based also on the least distance rule. The CBS-AN link is used for control signalling and it is not involved in the data path. A backhaul link (not shown in Fig. 5.2) may be required between the CBS and the DBS for signalling and coordination.
purposes. As with the CBS-AN link, the DBS-CBS link is not involved in data transmission.

5.3 Handover-related Core-network Signalling Load

This section builds upon the system model in Fig. 5.2 and derives the CN signalling load in both HO schemes. Assume that each inter-CBS HO in the CN-transparent HO generates CN signalling $S_1$, while intra-CBS HOs do not generate CN signalling as discussed earlier. On the other hand, each HO in the CN-visible HO generates CN signalling $S_2$. Notice that $S_1$ and $S_2$ represent the signalling load towards the CN generated by a single inter-CBS HO (in the CDSA with CN-transparent HO) and a single conventional CN-visible HO, 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 CN signalling overhead. In this model, $S_1$ and $S_2$ are referred to as CN signalling load irrespective of the measured quantity.

5.3.1 Core-network Signalling in CN-transparent Handover

The total CN signalling load $S_1$ generated by an active UE during the life time of a session depends on the session duration\(^1\) distribution, the UE mobility\(^2\) and the BS density\(^3\). The expected value of the CN signalling load generated by a UE in the CDSA with CN-transparent HO $\mathbb{E}[S_1]$ can be calculated as:

$$
\mathbb{E}[S_1] = \sum_{i=0}^{\infty} s_{1,i} \cdot f_{S_1}(s_{1,i}),
$$

where $s_{1,i} = i \cdot S_1$ and $f_{S_1}(s_{1,i})$ is the probability that the CN signalling load in the CN-transparent HO scheme is $s_{1,i}$. This probability can be calculated by using the Markov Chain shown in Fig. 5.3. Here, $P_1^0$ refers to the probability that the UE will not generate

\(^1\) 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.

\(^2\) The term “cell residence time” is used to model the UE mobility. It is defined as the total time spent by a UE in a single cell.

\(^3\) Assuming that the transmit power is the same for all cells in the same tier.
Figure 5.3: Markov Chain of the handover related core-network signalling

CN signalling in the CN-transparent HO scheme, and $P_1^g$ is the probability that the UE will not generate more CN signalling given that it has already generated CN signalling. Notice that $P_1^o \leq P_1^g$ because the residual session duration at an arbitrary time instance is typically less than (or equals to) the residual session duration at any previous time instance. Since the amount of signalling generated by the UE increases with the time, a transition from state $s_{1,j}$ to state $s_{1,i}$ has a zero probability when $j > i$. Based on this model, $f_{s_1}(s_{1,i})$ can be formulated as:

$$
\begin{align*}
  f_{s_1}(s_{1,i}) &= \begin{cases} 
    P_1^o, & \text{for } i = 0, \\
    Q_1 P_1^o (1 - P_1^o) (1 - Q_1 P_1^o)^{i-1}, & \text{for } i \geq 1,
  \end{cases}
\end{align*}
$$

(5.2)

where $Q_1 \geq 1$ is the ratio between $P_1^g$ and $P_1^o$. The CDF of the CN signalling $F_{s_1}(s_{1,i})$ can be written as:

$$
F_{s_1}(s_{1,i}) = \sum_{j=0}^{i} f_{s_1}(s_{1,j}) = P_1^o - \left((1 - P_1^o) \left((1 - Q_1 P_1^o)^i - 1\right)\right).
$$

(5.3)

Substituting (5.2) into (5.1) and simplifying the resultant equation gives the expected CN signalling in the CDSA with CN-transparent HO as a function of $P_1^o$, $Q_1$ and $S_1$:

$$
E[S_1] = \frac{S_1}{Q_1} \left(\frac{1}{P_1^o} - 1\right).
$$

(5.4)

Fig. 5.4 shows the effect of $Q_1$ on the CDF of the CN signalling load, i.e., $F_{s_1}(s_{1,i})$. It can be seen that as $Q_1$ increases, the probability of generating large amount of CN signalling decreases while the probability of zero CN signalling load remains constant. In other words, the transition probability from state $s_{1,0}$ to state $s_{1,1}$ increases as $Q_1$ increases, while the transition probability from state $s_{1,i}$ to state $s_{1,j}$ decreases, where $i \geq 1$ and $j > i$. This result indicates that $P_1^o$ and $Q_1$ are important parameters that have a significant influence on the total HO-related CN signalling generated by the active UE. In the following, these parameters are derived first under general distributions.
5.3. Handover-related Core-network Signalling Load

Figure 5.4: CDF of the normalised CN signalling load in the CN-transparent handover scheme, with $P_o = 0.25$

for the session duration and the cell residence time, and then closed-form expressions are obtained in a special case where the session duration and the cell residence time are exponentially distributed.

General distribution for session duration and cell residence time

Consider the CDSA system model described in Section 5.2 and Fig. 5.2, where $\lambda_d > \lambda_c$. Notice that in typical networks $\lambda_d \gg \lambda_c$. Assume a session duration $D$ with probability density function (PDF) $f_D(d)$ and mean $E[D]$. The CBS residence time is modelled as a random variable $R_c$ with PDF $f_{R_c}(r_c)$ and mean $E[R_c]$, while the DBS residence time is modelled as a random variable $R_d$ with PDF $f_{R_d}(r_d)$ and mean $E[R_d]$. Fig. 5.5 provides a timing diagram that illustrates the definition of these parameters, without loss of generalisation. In addition, assume that the users move at random directions with random velocities. Notice that this assumption yields the worst-case performance. Under this assumption, $E[R_c]$ ($E[R_d]$) 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 [147]. In other words, $E[R_c]$ ($E[R_d]$) can be considered as the inverse of the CBS (DBS) HO rate. Following the derivations in [147], the average number of UE in
5.3. Handover-related Core-network Signalling Load

A CBS $N_{c1}$ can be calculated by:

$$N_{c1} = U A_c,$$  \hspace{1cm} (5.5)

where $U$ is the UE density and $A_c$ is the average CBS area. On the other hand, the average number of UE leaving a CBS per unit of time $N_{c2}$ can be obtained as the CBS boundary crossing rate, i.e.,

$$N_{c2} = \frac{U \mathbb{E}[V] \varsigma_c}{\pi},$$  \hspace{1cm} (5.6)

where $\varsigma_c$ is the average CBS perimeter, while $\mathbb{E}[V]$ is the average user speed. $\mathbb{E}[R_c]$ can be obtained by dividing (5.5) by (5.6), i.e.,

$$\mathbb{E}[R_c] \approx \frac{\pi A_c}{\mathbb{E}[V] \varsigma_c},$$  \hspace{1cm} (5.7)

Considering the PPP model, i.e., $A_c \approx \frac{1}{\lambda_c}$ and $\varsigma_c \approx \frac{4}{\sqrt{\lambda_c}}$ [148], then $\mathbb{E}[R_c]$ can be rewritten as:

$$\mathbb{E}[R_c] \approx \frac{\pi}{4 \mathbb{E}[V] \sqrt{\lambda_c}}.$$  \hspace{1cm} (5.8)

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

$$\mathbb{E}[R_d] \approx \frac{\pi}{4 \mathbb{E}[V] \sqrt{\lambda_d}}.$$  \hspace{1cm} (5.9)
5.3. Handover-related Core-network Signalling Load

Consider a UE associated with CBS\(_i\) and DBS\(_m\). From the CN-transparent HO model in Fig. 5.2a, it can be noticed that the CN signalling is generated in inter-CBS HOs only. Expressed differently, all the DBS HOs do not generate CN signalling as long as the CBS/DP anchor point remains the same. Thus the definition of \( P^o_1 \) is equivalent to the probability that the UE does not change CBS\(_i\) during the life time of the session, irrespective of the DBS HOs. In other words, \( P^o_1 \) is equivalent to the probability that the session duration is less than the residual residence time in CBS\(_i\), i.e.,

\[
P^o_1 = \text{Prob} \left[ D < R_{c,r} \right] = \int_0^\infty \int_x^y \frac{1}{E[R_c]} \int_y^\infty f_{R_c}(x) \, dx \, dy,
\]

where Prob \([\cdot]\) means probability of an event and \( R_{c,r} \) is the residual residence time in the CBS as shown in Fig. 5.5 with PDF \( f_{R_{c,r}}(y) \). The latter can be formulated as a function of the CBS residence time distribution based on the renewal theorem [149], [150]. The PDF of the residual CBS residence time can be expressed as:

\[
f_{R_{c,r}}(y) = \frac{1}{E[R_c]} \int_y^\infty f_{R_c}(x) \, dx.
\]

Taking the Laplace transform of (5.11) yields:

\[
\mathcal{L} \left\{ f_{R_{c,r}}(y) \right\} = \mathcal{L} \left\{ \frac{1}{E[R_c]} \int_y^\infty f_{R_c}(x) \, dx \right\},
\]

where \( \mathcal{L} \) is the Laplace transform operator. Alternatively, (5.12) can be written as:

\[
\mathcal{L} \left\{ f_{R_{c,r}}(y) \right\} = \frac{1}{E[R_c]} \left( \mathcal{L} \left\{ \int_0^\infty f_{R_c}(x) \, dx \right\} - \mathcal{L} \left\{ \int_0^y f_{R_c}(x) \, dx \right\} \right)
\]

\[
= \frac{1}{E[R_c]} \left( \mathcal{L} \left\{ 1 \right\} - \mathcal{L} \left\{ \int_0^y f_{R_c}(x) \, dx \right\} \right)
\]

\[
= \frac{1}{E[R_c]} \left( \frac{1}{s} - \frac{\mathcal{L} \left\{ f_{R_c}(y) \right\}}{s} \right).
\]

By taking the inverse Laplace transform of (5.13), the PDF of the CBS residence time can be obtained as:

\[
f_{R_{c,r}}(y) = \mathcal{L}^{-1} \left\{ \frac{1 - \mathcal{L} \left\{ f_{R_c}(y) \right\}}{s E[R_c]} \right\} = \mathcal{L}^{-1} \left\{ \frac{4 \, E[V] \sqrt{\lambda_c}}{s \pi} \left( 1 - \mathcal{L} \left\{ f_{R_c}(y) \right\} \right) \right\},
\]

where \( \mathcal{L}^{-1} \) is the inverse Laplace transform operator.

On the other hand, the definition of \( P^g_1 \) 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 \(P_1^g\) is equivalent to the probability that the residual session duration \(D_r\) is less than the CBS residence time, i.e.,

\[
P_1^g = \operatorname{Prob} [D_r < R_c] = \int_{z=0}^{\infty} \int_{u=0}^{z} f_{D_r}(u) \, du \, dz,
\]

(5.15)

where \(f_{D_r}(u)\) is the PDF of \(D_r\) which can be calculated by following the same approach of (5.11)–(5.14), i.e.,

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

(5.16)

Finally, \(Q_1\) can be calculated as the ratio between (5.15) and (5.10). Substituting the resultant \(Q_1\) and (5.10) in (5.4) gives:

\[
\mathbb{E}[S_1] = \frac{S_1 \left( 1 - \int_{y=0}^{\infty} \int_{x=0}^{y} f_{D}(x) \, dx \, dy \right)}{\int_{z=0}^{\infty} \int_{u=0}^{z} f_{D_r}(u) \, du \, dz}.
\]

(5.17)

**Exponential distribution for session duration and cell residence time**

This section considers the scenario where the session duration and the cell residence time are exponentially distributed as considered in [133, 151–154], such that

\[
f_D(u) = \frac{e^{-u/\mathbb{E}[D]}}{\mathbb{E}[D]},
\]

(5.18)

and

\[
f_{R_c}(y) = \frac{e^{-y/\mathbb{E}[R_c]}}{\mathbb{E}[R_c]} = \frac{4 \mathbb{E}[V] \sqrt{\lambda_c e^{-4 \mathbb{E}[V] \sqrt{\lambda_c y}}}}{\pi}.
\]

(5.19)

**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 (5.18) in (5.16) and simplifying the resultant equation yields \(f_{D_r}(u)\) in the same form as (5.18). Similarly, substituting (5.19) in (5.14) results in \(f_{R_{c,r}}(y)\) in the same form as (5.19). \(\square\)
5.3. Handover-related Core-network Signalling Load

The probability that the session duration is less than the residual CBS residence time, i.e., $P_1^o$, can then be obtained as

$$P_1^o = \frac{\pi}{4\mathbb{E}[V]\mathbb{E}[D]\sqrt{\lambda_c} + \pi}, \quad Q_1 = 1,$$

by substituting these values into (5.10) and (5.15) and solving the integrals. Finally, the expected value of the CN signalling load in the CDSA with CN-transparent HO under exponential distribution can be simplified by substituting (5.18), (5.19) and the results of Lemma 1 into (5.17), i.e.,

$$\mathbb{E}[S_1] = \frac{4}{\pi}S_1\mathbb{E}[V]\mathbb{E}[D]\sqrt{\lambda_c}.$$

5.3.2 Core-network Signalling in Conventional CN-visible Handover

The modelling approach proposed in Section 5.3.1 can be adapted to model the conventional CN-visible HO signalling in order to assess the CDSA gains. The expected CN signalling in the conventional CN-visible HO $\mathbb{E}[S_2]$ can be written as:

$$\mathbb{E}[S_2] = \frac{S_2}{Q_2}\left(\frac{1}{P_2^o} - 1\right),$$

where $P_2^o$ and $Q_2 = \frac{P_2^o}{P_2}$ have the same definitions as $P_1^o$ and $Q_1$ respectively, with the subscript 2 meaning parameters of the conventional CN-visible HO model. It should be noticed that each HO in the conventional model generates signalling towards the CN because each DBS becomes a DP anchor point. Thus, parameters of the MC layer (i.e., the CBS layer in the CDSA terminology) do not capture all the CN signalling load in the conventional CN-visible HO model. A more convenient design approach is to consider parameters of the SC layer (i.e., the DBS layer in the CDSA terminology), since for the dense deployment considered, the DBS HOs $\gg$ CBS HOs. As a result, $P_2^o$, $P_2^g$ and $Q_2$ can be calculated by following a similar approach as the one used in deriving equations (5.10)–(5.17), by replacing $R_{c,r}$ with $R_{d,r}$, $f_{R_{c,r}}(y)$ with $f_{R_{d,r}}(y)$, $f_{R_c}(y)$ with $f_{R_d}(y)$, and $\lambda_c$ with $\lambda_d$. The expected CN signalling load in the conventional CN-visible HO can now be formulated as:

$$\mathbb{E}[S_2] = \frac{S_2}{Q_2}\left(\frac{1}{P_2^o} - 1\right)\left(\frac{1}{\int_{y=0}^{\infty} f_{R_{d,r}}(y) dy} \int_{x=0}^{\infty} f_D(x) dx dy \right).$$

(5.23)
Similarly, when the session duration and the DBS residence time are exponentially distributed, it can be proved that parameters of the conventional CN-visible HO simplify to

\[ P_2^o = \frac{\pi}{4E[V]E[D]} \sqrt{\lambda_d + \pi}, \quad Q_2 = 1, \quad (5.24) \]

\[ E[S_2] = \frac{4}{\pi} S_2 E[V] E[D] \sqrt{\lambda_d}. \quad (5.25) \]

It can be noticed that the system model of both the CN-transparent and the CN-visible HOs becomes memoryless under exponential distribution, since \( Q_1 = Q_2 = 1 \) as depicted by (5.20) and (5.24). Expressed differently, \( P_1^o \) and \( P_2^o \) are independent of the previous state and they are equal to \( P_1^o \) and \( P_2^o \), respectively. From a signalling load perspective, this can be considered as the worst-case as shown in Fig. 5.4, thus an appropriate setting of network parameters becomes of great importance in this scenario.

The proportional relationship in (5.25) between \( E[S_2] \) and the DP anchor density in the conventional CN-visible HO suggests reducing the latter to minimise the CN signalling load. This can be achieved by moving the DP anchor point to the CBS of the CDSA (i.e., moving towards a CN-transparent HO) in order to exploit the lower density of the CBS since \( \lambda_c \ll \lambda_d \). The CDSA-based CN-transparent HO gain \( G \) in terms of CN signalling load reduction w.r.t. the conventional CN-visible HO can be obtained by:

\[ G = 1 - \frac{E[S_1]}{E[S_2]} \quad (5.26) \]

For the case of \( Q_1 = Q_2 = 1 \), the CN signalling reduction gain can be obtained by substituting (5.21) and (5.25) into (5.26), i.e.,

\[ G = 1 - \frac{S_1}{S_2} \sqrt{\frac{\lambda_c}{\lambda_d}}, \quad \text{with} \quad Q_1 = Q_2 = 1. \quad (5.27) \]

5.4 Data Plane Backhaul Latency

Exploiting the CBS as a DP anchor point could reduce the CN signalling overhead. Nonetheless, this gain could come at the expense of increasing the DP latency, which consists of backhaul delay and RAN delay. The former is the delay experienced by the packets from the CN to the DBS. On the other hand, the RAN delay is associated with
the air interface and it represents the DBS-UE delay. This section focuses only on the backhaul latency since the RAN latency is the same under both CDSA separation/HO schemes, and it does not depend on the HO model but rather on frame structure and access procedure. The DP-BL $L$ consists of propagation delay $L_p$, transmission delay $L_t$ and processing delay $L_s$. For the CN-transparent HO model, the DP traverses two links $AN \rightarrow CBS \rightarrow DBS$ as shown in Fig. 5.2a. On the other hand, the DP of the CN-visible HO case traverses a single link $AN \rightarrow DBS$ as in Fig. 5.2b. Assuming error-free transmission across the backhaul network, each latency component is modelled as in the following.

### 5.4.1 Data Plane Backhaul Propagation Delay

The DP propagation delay in the CN-visible HO network can be calculated by the following lemma.

**Lemma 2.** The propagation delay between an arbitrary AN $a$ in $\Phi_a$ and an arbitrary DBS $d$ in $\Phi_d$ located within the voronoi cell of $a$ is a Rayleigh distributed random variable with PDF:

$$f_{L_p}(l_p) = \begin{cases} 0, & \text{for } l_p < 0 \\ \frac{l_p}{\sigma_{a,d}} e^{-\frac{l_p^2}{2\sigma_{a,d}^2}}, & \text{for } l_p \geq 0 \end{cases}, \quad (5.28)$$

and expected value:

$$E[L_p] = \frac{1}{2\varphi_{a,d} \sqrt{\lambda_a}}, \quad (5.29)$$

where $\sigma_{a,d} = \frac{1}{\varphi_{a,d} \sqrt{2\pi \lambda_a}}$ is the scale parameter, and $\varphi_{a,d}$ is the propagation speed of the $a-d$ link.

**Proof.** The probability that the distance $R$ between $a$ in $\Phi_a$ and $d$ in $\Phi_d$ is greater than a distance $r_0$ is equivalent to the null probability in an area with radius $r_0$ centred at $d$, i.e.,

$$\text{Prob}[R > r_0] = e^{-\pi \lambda_a r_0^2}, \quad (5.30)$$
the CDF of $\mathcal{R}$ is $F_\mathcal{R}(r) = \text{Prob}[\mathcal{R} \leq r_0] = 1 - e^{-\pi \lambda_a r_0^2}$ and the PDF of $\mathcal{R}$ can be obtained as:

$$f_\mathcal{R}(r) = \frac{d}{dr} \left(1 - e^{-\pi \lambda_a r^2}\right) = 2\pi \lambda_a r e^{-\lambda_a \pi r^2}. \quad (5.31)$$

Since $L_p$ is a monotonic increasing function of $\mathcal{R}$ (i.e., $L_p$ is the ratio between $\mathcal{R}$ and $\phi_{a,d}$), then $F_\mathcal{R}(r)$ can be written as:

$$F_\mathcal{R}(r) = \int_{m_1}^{m_2} f_\mathcal{R}(r) \, dr = \int_{l_p(m_1)}^{l_p(m_2)} f_\mathcal{R}(r(l_p)) \, \frac{dr}{dl_p} \, dl_p. \quad (5.32)$$

Expressed differently, $f_{L_p}(l_p)$ can be derived from $f_\mathcal{R}(r)$ as:

$$f_{L_p}(l_p) = f_\mathcal{R}(r) \left| \frac{dr}{dl_p} \right|. \quad (5.33)$$

Substituting (5.31) and $r = l_p \phi_{a,d}$ in (5.33) and solving the resultant equation yields (5.28). The expected value of the propagation delay in (5.29) can be directly obtained by substituting (5.28) in $E[L_p] = \int_0^\infty l_p f_{L_p}(l_p) \, dl_p$.

Lemma 2 can also be used to calculate the propagation delay across a single link in the CN-transparent HO model. However, the DP path of the latter traverses two links. Thus the total propagation delay is a summation of Rayleigh variables. By using the method of [155], the PDF of the total propagation delay in the CN-transparent HO scheme can be expressed as

$$f_{L_p}(l_p) = \frac{\sigma_{a,c}^2 l_p e^{\frac{-l_p^2}{2(\sigma_{a,c}^2 + \sigma_{c,d}^2)}}}{\sigma_{a,c}^2 + \sigma_{c,d}^2} + \frac{\sigma_{c,d}^2 l_p e^{\frac{-l_p^2}{2(\sigma_{a,c}^2 + \sigma_{c,d}^2)}}}{\sigma_{a,c}^2 + \sigma_{c,d}^2} + \frac{\pi \mathcal{M} \sigma_{a,c} \sigma_{c,d} \left[l_p^2 - \sigma_{a,c}^2 - \sigma_{c,d}^2\right] e^{\frac{-l_p^2}{2(\sigma_{a,c}^2 + \sigma_{c,d}^2)}}}{2 \left(\sigma_{a,c}^2 + \sigma_{c,d}^2\right)^{\frac{3}{2}}}, \quad (5.34)$$

with

$$\mathcal{M} = \text{erf} \left( \frac{\sigma_{c,d} l_p}{\sigma_{a,c} \sqrt{2 \left(\sigma_{a,c}^2 + \sigma_{c,d}^2\right)}} \right) + \text{erf} \left( \frac{\sigma_{a,c} l_p}{\sigma_{c,d} \sqrt{2 \left(\sigma_{a,c}^2 + \sigma_{c,d}^2\right)}} \right), \quad (5.35)$$

where $\text{erf} \{u\} = \frac{2}{\sqrt{\pi}} \int_0^u e^{-v^2} \, dv$ is the error function. The expected value of the total propagation delay can now be written as:

$$E[L_p] = \int_0^\infty l_p f_{L_p}(l_p) \, dl_p = \sum_{\forall x \in \Lambda \setminus \{d\}} \frac{1}{2 \varphi_{x,y} \sqrt{\lambda_x}}, \quad (5.36)$$
where $\Lambda$ is the set of all nodes in the DP backhaul path, i.e., $\Lambda = \{a, d\}$ in the CN-visible HO scheme and $\Lambda = \{a, c, d\}$ in the CN-transparent HO scheme. $y$ is the node in the layer immediately below the layer of node $x$.

### 5.4.2 Data Plane Backhaul Transmission Delay

The transmission delay can be modelled as the ratio between the data load and the backhaul capacity. Depending on the backhaul technology (e.g., copper, fibre, wireless, etc.), repeater(s) may be needed if the link transmission range is smaller than the distance between the nodes. These repeaters result in additional transmission delays. The expected number of repeaters between node $x$ in $\Phi_x$ and node $y$ in $\Phi_y$, i.e., $E[\varrho_{x,y}]$, is:

$$E[\varrho_{x,y}] = \left\lceil E[R_{x,y}] \varpi_{x,y} - 1 \right\rceil = \left\lceil \frac{1}{2 \varpi_{x,y} \sqrt{\lambda_x}} - 1 \right\rceil,$$

(5.37)

where $E[R_{x,y}] = \int_{0}^{\infty} r f_R(r) \, dr$ is the expected value of the distance between nodes $x$ and $y$. $\varpi_{x,y}$ is the transmission range of the backhaul link between $x$ and $y$, and $[m]$ means the smallest integer $\geq m$, i.e., ceil operator. The transmission delay between $x$ and $y$ is $(\frac{Z}{\varpi_{x,y}}) \left( E[\varrho_{x,y}] + 1 \right)$, which consists of delays at $x$ and at each repeater, with $Z$ being the data load (i.e., packet size including overheads), and $\varpi_{x,y}$ is the capacity of the link connecting $x$ and $y$ through the repeaters. The expected value of the DP backhaul transmission delay can be written as:

$$E[L_t] = \sum_{\forall x \in \Lambda \setminus \{d\}} \frac{Z}{\varpi_{x,y}} \left\lceil \frac{1}{2 \varpi_{x,y} \sqrt{\lambda_x}} \right\rceil,$$

(5.38)

where $x$ and $y$ have the same definitions as in (5.36). Notice that the transmission delay modelling in this section is applicable to wired links.

### 5.4.3 Data Plane Backhaul Processing Delay

To model the DP backhaul processing delay, the approach of [146] and [156] has been followed by assuming that $L_s$ in a single node has a gamma distribution, since the latter has been found to provide a good fit for real measurements [156], [157]. The processing delay in a single node $x$ depends on: processing capabilities of the node, DBS data load
and the number of DBSs connected to the node. The latter can be captured by the shape parameter \( \kappa \) of the gamma distribution, i.e., \( \kappa_x \propto \frac{\lambda_y}{\alpha_x} \), while the scale parameter \( \theta_x \) captures the processing capabilities and the DBS data load. Based on the classical definition of the gamma distribution [158], the PDF of node \( x \) processing delay is:

\[
f_{L_x}(l_x) \sim \Gamma(\kappa_x, \theta_x) = \frac{l_x^{\kappa_x-1} e^{-\frac{l_x}{\theta_x}}}{\theta_x^{\kappa_x} \Gamma(\kappa_x)} = \frac{l_x^{\mu_x \lambda_x d - 1} e^{-\frac{l_x}{\alpha_x + \beta_x z}}}{(\alpha_x + \beta_x Z)^{\mu_x \lambda_x d} \Gamma(\mu_x \lambda_x d)},
\]

where \( \kappa_x = \frac{\mu_x \lambda_y}{\lambda_x} \) with \( \mu_x \) being the constant of proportionality, \( \theta_x = \alpha_x + \beta_x Z \) with \( \alpha_x \) being the static processing delay and \( \beta_x \) is the processing delay per bit. \( \Gamma(U) \) is the gamma function evaluated at \( U \). The processing delay at the AN, the CBS or the DBS can be modelled by (5.39). For a tandem topology based repeaters, the amount of DP traffic traversing a repeater between \( x \) and \( y \), where \( \lambda_y > \lambda_x \) and \( y \) is in the voronoi cell of \( x \), is the same amount of DP traffic traversing \( y \) because there could be more than a node of \( \Phi_y \) connected to \( x \). Thus, the shape parameter \( \kappa_{x,y} \) of the processing delay of all repeaters between \( x \) and \( y \) is proportional to \( \lambda_y \mathbb{E}[\varrho_{x,y}] \).

Assume that the processing capabilities of the nodes in the same tier is the same (e.g., all ANs have the same processing capabilities). However, nodes in different tiers may have different processing capabilities (e.g., nodes towards the CN may have higher processing capabilities than nodes towards the edge). As a result, the distribution of the total processing delay of all nodes and repeaters is not a simple gamma distribution because the scale parameter \( \theta \) of different nodes is not constant. The resultant total processing delay can be represented by a single gamma-series with coefficients that are computed by recursive relations. Theorem 1 [158] provides the PDF of this gamma-series.

**Theorem 1.** The sum \( G \) of \( \tilde{n} \) independent gamma variables \( \Gamma(\kappa_1, \theta_1), \Gamma(\kappa_2, \theta_2), \ldots, \Gamma(\kappa_{\tilde{n}}, \theta_{\tilde{n}}) \) can be represented by a single gamma-series with PDF:

\[
f_G(g) = \prod_{i=1}^{\tilde{n}} \left( \frac{\theta_{m_i}}{\theta_i} \right)^{\kappa_i} \sum_{q=0}^{\infty} \frac{g^{\rho+q-1} e^{-g_m}}{\Gamma(\rho+q) \theta_m^{\rho+q}},
\]

(5.40)
with:

\[ \rho = \sum_{i=1}^{\hat{n}} \kappa_i , \quad \theta_m = \min_{\forall i} \theta_i , \quad \Xi_0 = 1 \]

\[ \Xi_{q+1} = \frac{1}{q+1} \sum_{i=1}^{q+1} i \vartheta_i \Xi_{q+1-i} , \quad \vartheta_i = \sum_{j=1}^{\hat{n}} \omega_j \left( 1 - \frac{\theta_m}{\theta_j} \right) i. \]

**Proof.** The reader is referred to [158] for the proof of this theorem.

Finally, the expected value of the DP backhaul processing delay can be computed as

\[
\mathbb{E}[L_s] = \int_0^\infty l_s f_{L_s}(l_s) dl_s = \sum_{i=1}^{\hat{n}} \kappa_i \theta_i \\
= \sum_{\forall x \in \Lambda \setminus \{d\}} \mu_{x,y} \lambda_x \frac{1}{2 \varphi_{x,y} \sqrt{\lambda_x}} \left( \frac{1}{2 \varphi_{x,y} \sqrt{\lambda_x}} - 1 \right) \left( \alpha_{x,y} + \beta_{x,y} Z \right) + \sum_{\forall x \in \Lambda} \mu_x \lambda_x \left( \frac{1}{2 \varphi_{x,y} \sqrt{\lambda_x}} - 1 \right) \left( \frac{Z}{\varphi_{x,y}} + \Psi_{x,y} \right),
\]

(5.41)

where \( \mu_{x,y}, \alpha_{x,y} \) and \( \beta_{x,y} \) are the processing parameters of the repeaters between nodes \( x \) and \( y \).

### 5.4.4 Upper Bound of DBS Density Under Latency Constraints

The expected value of the DP-BL can be written as the summation of (5.36), (5.38) and (5.41). The exact closed-form expression is formulated as

\[
\mathbb{E}[L] = \sum_{\forall x \in \Lambda} \mu_x \lambda_x \left( \alpha_x + \beta_x Z \right) + \sum_{\forall x \in \Lambda \setminus \{d\}} \frac{1}{2 \varphi_{x,y} \sqrt{\lambda_x}} \left( \frac{1}{2 \varphi_{x,y} \sqrt{\lambda_x}} - 1 \right) \left( \alpha_{x,y} + \beta_{x,y} Z \right) + \sum_{\forall x \in \Lambda} \mu_x \lambda_x \left( \frac{1}{2 \varphi_{x,y} \sqrt{\lambda_x}} - 1 \right) \left( \frac{Z}{\varphi_{x,y}} + \Psi_{x,y} \right),
\]

(5.42)

where \( \Psi_{x,y} = \frac{\mu_{x,y} \lambda_x (\alpha_{x,y} + \beta_{x,y} Z)}{\lambda_y} \). A constrained DBS deployment is proposed in this section based on the DP-BL. Assuming a maximum tolerable DP-BL of \( \mathbb{E}[L_{thr}] \), then a DP latency protection condition can be formulated as:

\[
\mathbb{E}[L] \leq \mathbb{E}[L_{thr}].
\]

(5.43)

Based on this condition, the upper bound of the DBS density \( \lambda_{du} \) without violating (5.43) can be obtained by substituting (5.42) in (5.43) and solving the resultant equation
5.5 Numerical and Simulation Results

for $\lambda_d$. The exact closed-form expression of $\lambda_{du}$ is given by

$$\lambda_{du} \leq \frac{\mathbb{E}[L_{thr}] - B - \mu_{t,d}}{\mu_{a} (\alpha + \beta_a Z) + \mu_{c} (\alpha + \beta_c Z) + \mu_{a,c} \left( \frac{1}{2 \omega_{x_d} \sqrt{\lambda_x}} - 1 \right)} (\alpha_{i,d} + \beta_{i,d} Z) - \mu_d (\alpha_d + \beta_d Z)^{(5.44)}$$

with

$$B = \sum_{x \in \Lambda \setminus \{d\}} \frac{1}{2 \varphi_{x,y} \sqrt{\lambda_x}} + \frac{Z}{\varphi_{x,y}} \left[ \frac{1}{2 \omega_{x,y} \sqrt{\lambda_x}} - 1 \right] (\alpha_{a,c} + \beta_{a,c} Z)^{(5.45)}$$

where $i = c$ and $j = 1$ for the CN-transparent HO signalling scheme, while $i = a$ and $j = 0$ for the CN-visible HO signalling model.

### 5.5 Numerical and Simulation Results

This section presents and discusses numerical and simulation results that assess superiority of the proposed CN-transparent HO signalling scheme (under CDSA) over the conventional CN-visible HO signalling approach (under both CDSA and conventional RAN architecture). First, numerical and simulation results based on the analytical model in Section 5.3 are provided, where the CN signalling load of both schemes is compared. Then, the densification limit under DP-BL constraints is analysed and the impact of several parameters is evaluated. Finally, the CN-transparent HO signalling scheme is integrated with the predictive HO signalling models of Chapter 4, and simulation results are presented to evaluate the overall gains of the integrated scheme in terms of HO latency.

#### 5.5.1 Core-network Signalling Load Results

Here, the CN signalling load of both HO schemes is evaluated, where the session duration and the cell residence time are exponentially distributed. For a fair comparison, the CN signalling load (in terms of expected value and CDF) is normalised with $S_1$ in the CN-transparent HO scheme, and with $S_2$ in the conventional CN-visible HO scheme. Fig. 5.6 shows the probability of not generating CN signalling vs the user speed while Fig. 5.7 provides this probability vs the mean session duration for several
5.5. Numerical and Simulation Results

![Figure 5.6](image)  
**Figure 5.6:** Probability of not generating CN signalling vs user speed, $E[D] = 5$ min and $S_1 = S_2$

![Figure 5.7](image)  
**Figure 5.7:** Probability of not generating CN signalling vs session duration, with $E[V] = 30$ km/hr and $S_1 = S_2$

DBS densities with $S_1 = S_2$. As can be seen in Fig. 5.6, $P_1^o$ and $P_2^o$ decrease as the user speed increases, since a higher speed results into a higher inter-CBS HO probability. However, the CDSA with CN-transparent HO provides 4–6 times higher zero CN signalling probability as compared with the conventional CN-visible HO, and the CDSA improvement increases as the DBS density increases. Similarly to the speed effect, Fig. 5.7 indicates that the probability of not generating signalling towards the CN decreases as the session duration increases. This can be traced to the fact that
the HO happens in active state only. Hence, a longer session duration results in a higher activity probability which itself leads to a higher inter-CBS HO probability. The improvement from the CN-transparent HO scheme can be seen in Fig. 5.7 where \( P_1^0 \approx 3 P_2^0 \) with \( \lambda_d = 10 \lambda_c \) and \( \mathbb{E}[D] > 6 \text{ min} \). With ultra-high DBS densities, e.g., \( \lambda_d = 100 \lambda_c \), the CN-transparent HO improvement reaches 900% (i.e., \( P_1^0 \approx 10 P_2^0 \)) when \( \mathbb{E}[D] \geq 10 \text{ min} \).

Fig. 5.8a shows the normalised expected value of the CN signalling load (i.e., \( \frac{\mathbb{E}[S_1]}{S_1} \) and \( \frac{\mathbb{E}[S_2]}{S_2} \)) vs \( \mathbb{E}[V] \) while Fig. 5.8b provides the CDF of the normalised CN signalling load for low speed (3 km/hr), medium speed (30 km/hr) and high speed (100 km/hr) users, with \( \mathbb{E}[D] = 5 \text{ min} \), \( S_1 = S_2 \) and \( \lambda_d = 40 \lambda_c \). At the 90\(^{th}\) percentile, the CN signalling generated in the CN-transparent HO with low, medium and high user speed is \( \leq \{1, 7, 24\} \cdot S_2 \) respectively. For the same percentile, the CN signalling load in the conventional CN-visible HO is \( \leq \{5, 46, 155\} \cdot S_2 \) with low, medium and high user speed, respectively. Expressed differently, the CN-transparent HO signalling gain increases with the user speed. An interesting finding from Fig. 5.8 is that the CN signalling generated in the CN-transparent HO scheme with a medium user speed (i.e., 30 km/hr) is roughly the same as the signalling generated in the conventional CN-visible HO scheme with a low user speed (i.e., 3 km/hr). This can be linked to the CN-transparent HO system model and the proportional relationship in (5.21) between \( \mathbb{E}[S_1] \) and the term \( \mathbb{E}[V] \cdot \sqrt{\lambda_c} \), and in (5.25) between \( \mathbb{E}[S_2] \) and the term \( \mathbb{E}[V] \cdot \sqrt{\lambda_d} \), where \( \lambda_c \ll \lambda_d \). Thus it can be said that the CDSA with CN-transparent HO supports medium and high speed users with a significantly less CN signalling as compared with the conventional CN-visible HO scheme.

Fig. 5.9a shows the normalised expected value of the CN signalling load vs \( \mathbb{E}[D] \) while Fig. 5.9b provides the CDF of the normalised CN signalling load for several session durations, with \( \mathbb{E}[V] = 30 \text{ km/hr} \), \( S_1 = S_2 \) and \( \lambda_d = 40 \lambda_c \). As can be seen, the HO-related CN signalling in the conventional CN-visible HO scheme increases significantly as the session duration increases. Although the CN-transparent HO signalling load is also proportional to \( \mathbb{E}[D] \), the latter has a less effect on the CN-transparent HO signalling scheme as compared with the conventional CN-visible HO approach. Fig. 5.10 shows the effect of the DBS density on the CN signalling load with \( \mathbb{E}[V] = 30 \text{ km/hr} \),
5.5. Numerical and Simulation Results

![Graph](image)

(a) Normalised expected value of the CN signalling load

![Graph](image)

(b) CDF of the normalised CN signalling load with $\lambda_d = 40 \lambda_c$

Figure 5.8: User speed effect on the normalised CN signalling load, with $E[D] = 5$ min and $S_1 = S_2$

$E[D] = 5$ min and $S_1 = S_2$. It can be noticed that the CN signalling in the CN-visible HO increases as $\lambda_d$ increases. On the other hand, the CN signalling load of the CN-transparent HO scheme does not depend on the DBS density but rather it depends on the CBS density. With a 90% probability, the CN-transparent HO signalling load is $\leq 7S_2$, while the conventional CN-visible HO load is $\leq \{23, 45, 60, 75\} \cdot S_2$ with $\lambda_d = \{10, 40, 70, 100\} \cdot \lambda_c$ respectively.

System level simulations have been performed to validate the proposed modelling
5.5. Numerical and Simulation Results

![Graph](image)

(a) Normalised expected value of the CN signalling load

![Graph](image)

(b) CDF of the normalised CN signalling load with $\lambda_d = 40 \lambda_c$

Figure 5.9: Session duration effect on the normalised CN signalling load, with $S_1 = S_2$ and $\mathbb{E}[V] = 30$ km/hr

The considered network topology consists of one MME/S-GW, 19 omnidirectional DBSs and 1–4 CBSs. It has been assumed that the messages in the UE-CBS link are delivered correctly. In addition, a HO between DBSs under the control of different CBSs triggers a CBS HO. The DBS HO criteria follows the SS-based HO approach in (4.29). The HO procedure follows the signalling flow without prediction, as illustrated in Fig. 4.1 where the HO signalling messages 8–11 are replaced with an acknowledgement message in the CN-transparent HO scheme.
5.5. Numerical and Simulation Results

Figure 5.10: DBS density effect on the CDF of the normalised CN signalling load, with $E[V] = 30$ km/hr and $E[D] = 5$ min

Figure 5.11: CN signalling reduction in CN-transparent handover w.r.t. CN-visible handover, theoretical vs simulation

Other simulation parameters follow Table 4.4.

Fig. 5.11 compares the theoretical and the simulated gain in terms of CN signalling load reduction in the CN-transparent HO scheme w.r.t. the conventional CN-visible HO approach, while Fig. 5.12 shows the theoretical gain for other density and configuration values. As can be seen in Fig. 5.11, the gain values obtained from the simulation are in line with the theoretical values which validates the proposed modelling approach. When both schemes are used under CDSA configuration (i.e., $S_1 = S_2$), the CN-transparent
5.5. Numerical and Simulation Results

HO reduces the CN signalling by 70%–90% w.r.t. the conventional CN-visible HO, as shown in Fig. 5.12. This result indicates that the CN-transparent HO scheme is more beneficial (from the CN signalling load perspective) in dense deployment scenarios, since the gain increases with the DBS density.

It is worth mentioning that there is no standard procedure for the CDSA-based 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 RAN architecture. To compare the CDSA-based CN-transparent HO scheme with the conventional RAN architecture-based CN-visible HO approach, several cases are considered whilst assuming that $S_1 \geq S_2$. As can be seen in Fig. 5.12, the CDSA with CN-transparent HO provides 38%–80% signalling reduction gain when the inter-CBS HO procedure generates double the CN signalling load that is generated by a single HO in the conventional RAN architecture. Furthermore, a 5%–70% gain can be achieved when the inter-CBS HO procedure generates three times the CN signalling load generated by the conventional RAN HO procedure. Thus it can be concluded that in dense deployment scenarios the CDSA with CN-transparent HO can significantly reduce the overall signalling towards the CN, even if the inter-CBS HO procedure is more complicated than the conventional RAN HO procedure.

Figure 5.12: Theoretical CN signalling reduction in CN-transparent handover w.r.t. CN-visible handover
Table 5.1: Backhaul link and processing parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \varphi_{a,c} )</td>
<td>( 2.1 \times 10^8 ) m/s</td>
<td>( \mu_a )</td>
<td>1</td>
</tr>
<tr>
<td>( \varpi_{a,c} )</td>
<td>20 km</td>
<td>( \alpha_a )</td>
<td>1 ( \mu s )</td>
</tr>
<tr>
<td>( \varrho_{a,c} )</td>
<td>1 Gb/s</td>
<td>( \beta_a )</td>
<td>1 ns/bit</td>
</tr>
<tr>
<td>( \varphi_{c,d} )</td>
<td>( 2.1 \times 10^8 ) m/s</td>
<td>( \mu_c, \mu_{a,c} )</td>
<td>2</td>
</tr>
<tr>
<td>( \varpi_{c,d} )</td>
<td>20 km</td>
<td>( \alpha_c, \alpha_{a,c} )</td>
<td>2 ( \mu s )</td>
</tr>
<tr>
<td>( \varrho_{c,d} )</td>
<td>0.1 Gb/s</td>
<td>( \beta_c, \beta_{a,c} )</td>
<td>2 ns/bit</td>
</tr>
<tr>
<td>( \varphi_{a,d} )</td>
<td>( 2.1 \times 10^8 ) m/s</td>
<td>( \mu_d, \mu_{c,d}, \mu_{a,d} )</td>
<td>4</td>
</tr>
<tr>
<td>( \varpi_{a,d} )</td>
<td>20 km</td>
<td>( \alpha_d, \alpha_{c,d}, \alpha_{a,d} )</td>
<td>4 ( \mu s )</td>
</tr>
<tr>
<td>( \varrho_{a,d} )</td>
<td>0.1 Gb/s</td>
<td>( \beta_d, \beta_{c,d}, \beta_{a,d} )</td>
<td>4 ns/bit</td>
</tr>
</tbody>
</table>

5.5.2 DP-BL Constrained DBS Deployment Results

This section provides and discusses numerical results based on the proposed DP-BL model for both HO schemes. As an exemplary case, fibre optic backhaul links are considered with parameters and processing capabilities provided in Table 5.1. Fig. 5.13 shows the CBS and the DBS densities effect on the DP-BL, while Fig. 5.14 shows the effect of the CBS density and the maximum tolerable DP-BL on the upper bound of DBS density under DP-BL constrains. It can be noticed that relaxing \( \mathbb{E}[L_{thr}] \) increases the allowed DBS density\(^4\). This can be traced to the proportional relationship between the DBS density and the DP-BL. With a conservative latency threshold of \( \mathbb{E}[L_{thr}] < 1 \) ms, both HO schemes can operate with roughly the same DBS density without violating the latency constraint. Although increasing \( \mathbb{E}[L_{thr}] \) increases the allowed DBS density for both HO schemes, the rate of increase in the allowed density is higher in the CN-visible HO signalling scheme as compared with the CN-transparent HO signalling model. In other words, the former can operate with a higher DBS density

\(^4\)The terms “allowed DBS density” and “upper bound of DBS density under DP-BL constraints” are used interchangeably.
than the later (from the DP-BL perspective), as can be seen in Fig. 5.14.

This observation implies that the rate of increase in the DP-BL with the increase of the DBS density is higher in the CN-transparent HO signalling scheme as compared with the CN-visible HO signalling approach. Expressed differently, although the DP-BL of both HO schemes is proportional to the DBS density, the latter has a higher impact on the DP-BL of the CN-transparent HO signalling scheme as can be seen in Fig. 5.13. Nonetheless, Fig. 5.14b indicates that the CBS density has a significant influence on the allowed DBS density in the CN-transparent HO signalling scheme. In
the latter, the proportional relationship between the DBS density and the DP-BL has a large (small) slope with low (high) CBS densities, as shown in Fig. 5.13b. This can be linked to the fact that increasing the CBS density distributes the processing load (i.e., smaller number of DBSs associated with each CBS), which reduces the DP-BL and hence increases the allowed DBS density. Moreover, a higher CBS density results in a smaller CBS-DBS distance which reduces the propagation delay and alleviates the additional delay at intermediate repeaters. As a result, the CBS density is an important design parameter that can be controlled to reduce (to increase) the DP-BL (the upper
5.5. Numerical and Simulation Results

Figure 5.15: Effect of data load on the upper bound of DBS density under DB-BL constraints, with $\lambda_1 = 5$

bound of DBS density under DP-BL constraints) in the CN-transparent HO signalling scheme.

Fig. 5.15 shows the effect of the data load on the upper bound of DBS density under DP-BL constraints. It indicates that the amount of increase in the allowed DBS density remains roughly constant, in both HO schemes, irrespective of the data load when $Z > 2000$ bits (i.e., 250 bytes). Thus, it can be said that the rate of increase in the allowed DBS density is marginally affected by the data load and significantly influenced by the CBS density and the latency threshold. Based on the results of Sections 5.5.1 and
5.5.2, it can be concluded that the proposed CN-transparent HO signalling scheme is more beneficial (from the CN signalling overhead perspective) in dense DBS deployment scenarios. However, an appropriate setting of network parameters is of great importance to ensure that the gains in such scenarios do not come at the expense of violating the DP-BL constraints.

5.5.3 Integrated Handover Signalling Scheme Results

The predictive HO signalling models in Chapter 4 have been integrated with the CN-transparent HO signalling scheme, in order to evaluate the overall gains of the proposed integrated scheme over the conventional non-predictive and CN-visible HO signalling approach. The HO procedure in Fig. 4.1 has been adopted. In addition, the HO latency parameters follow Table 4.3. For the history-based predictive DBS HO signalling scheme, the simulation scenarios and parameters of Section 4.5.1 have been considered. On the other hand, the measurement-based context-aided predictive DBS HO signalling scheme has been evaluated under the simulation scenarios and parameters of Table 4.4. The evaluation is based on the HO latency reduction in the integrated scheme w.r.t. the conventional non-predictive and CN-visible HO approach. Fig. 5.16 shows the HO latency reduction in the integrated history-based predictive and CN-transparent HO, while Fig. 5.17 provides the HO latency reduction in the integrated measurement-based context-aided predictive and CN-transparent HO.

In the history-based prediction, the overall gains depend on the UE mobility profile as well as the prediction confidence requirements as shown in Fig. 5.16. Recalling the results of Section 4.5.1, the integrated scheme provides roughly double the gains of the predictive-only HO. For instance, the HO latency reduction gain increases from 34% to 60% in the regular mobility scenario with $h_{thr} \geq 0.3$ hartley, and from 23% to 49% in the 30% random mobility scenario with $h_{thr} \geq 0.6$ hartley. With a very conservative prediction confidence setting of $h_{thr} < 0.08$ hartley, the integrated scheme reduces the HO latency by 27%, which is significantly higher the 0% reduction in the predictive-only HO. Similarly, Fig. 5.17 indicates that the integrated scheme doubles the peak gains of the predictive-only HO, i.e., from 33.6% in Fig. 4.16 to 60% in Fig. 5.17.
5.5. Numerical and Simulation Results

Figure 5.16: Handover signalling latency reduction in the integrated history-based predictive and CN-transparent handover w.r.t. the conventional non-predictive and CN-visible handover.

Figure 5.17: Handover signalling latency reduction in the integrated measurement-based context-aided predictive and CN-transparent handover w.r.t. the conventional non-predictive and CN-visible handover.

At the regions with a poor prediction performance, e.g., low SS triggering threshold and no HO hysteresis, the integrated scheme yields more than 10-fold the gains of
the predictive-only HO. Overall, the integrated scheme with either history-based or measurement-based context-aided prediction reduces the HO latency by up to 60% as compared with the conventional non-predictive and CN-visible HO approach.

5.6 Conclusion

This chapter focuses on the CN signalling part of the HO process and proposes a CN-transparent HO scheme with minimal CN signalling overhead. The CDSA with dual connectivity is considered as a base architecture, and the CBS is utilised as a mobility anchor for the UE and as a DP anchor for the DBSs, resulting in a CN-transparent HO signalling model. An analytical framework has been developed to assess superiority of this scheme over the conventional HO signalling approach. Both generic and exponential distributions are considered for the session duration and the cell residence time, and closed-form expressions are obtained for the expected value of the CN signalling load as well as the probability of a fully CN-transparent HO signalling. In addition, the DP-BL has been identified as the main challenge of the proposed scheme. Thus, an analytical model has developed to investigate this issue based on stochastic geometry, and an upper bound for the DBS density under latency constraints has been derived.

It has been found that the proposed CN-transparent HO signalling scheme can significantly reduce the HO-related signalling towards the CN. The modelling approach resulted in a CN signalling load proportional to the user speed and the session duration. Nonetheless, the slope of this relationship in the CN-transparent HO is significantly less than the CN-visible HO slope due to the large CBS footprint. It has been observed that medium speed users who follow the CN-transparent HO generate roughly the same CN signalling as low speed users who follow the CN-visible HO. Moreover, the CN-transparent HO with high user speed generates a significantly less CN signalling than the CN-visible HO with medium or high user speed. In probabilistic terms, the CN-transparent HO achieves 600%−900% higher zero CN signalling probability over the CN-visible HO approach. In ultra-dense deployment scenarios, the CN-transparent HO achieves the peak gains which reach 90% reduction in the CN-signalling load w.r.t.
the conventional CN-visible HO approach.

From the DP latency perspective, it has been found that the DP-BL is proportional to the DBS density under both HO schemes. However, the slope of this relationship is inversely proportional to the CBS density in the CN-transparent HO model. This suggests a careful CBS/DBS deployment to exploit mobility benefits of the CDSA with CN-transparent HO signalling whilst minimising the DP-BL. In this direction, the proposed DP-BL constrained DBS deployment can play a key role. Finally, from the HO latency perspective, the CN-transparent HO signalling scheme provides roughly the same gains as the predictive HO models. When both schemes are integrated, up to 60% reduction in the HO latency is achieved w.r.t. the conventional non-predictive and CN-visible HO signalling approach.
Chapter 6

Epilogue

6.1 Thesis Summary and Conclusions

As soon as 2020, network densification will be the dominant theme to support enormous capacity and massive connectivity. However, such deployment scenarios raise several challenges and they impose new constraints. In particular, signalling load, mobility management and EE will become critical considerations in the 5G era. In this context, the conventional worst-case design approach may not be suitable because it wastes the transmission resources by over-provisioning the PL frame. This calls for the design of adaptive rather than static signalling mechanisms with dynamic frame allocations. In addition, the expected high HO rates in dense deployment scenarios could lead to a dramatical increase in mobility-related signalling overhead. From an energy perspective, the always-on service approach has been found to be the main source of inefficiency in the traditional system. These aspects suggest a paradigm shift towards a signalling and energy conscious RAN architecture with intelligent mobility management. In this direction, a new RAN architecture with logical separation of CP and DP was proposed to overcome limitations of the conventional architecture. The main concept of the CDSA depends on separating the coverage signals from those needed to support high data rate transmission. This allows data services to be provided by efficient DBSs under the umbrella of a coverage layer supported by CBSs. Such an approach opens a wide range of benefits with relaxed constraints.
6.1. Thesis Summary and Conclusions

In Chapter 2, a comprehensive survey of existing literature on the CDSA was presented. The concept, fundamentals and general structure were reviewed. In addition, the user states, i.e., detached, idle and active, were associated with the relevant CDSA layer(s). Based on this association, the functionalities of each plane were identified and mapped to each CDSA layer. Moreover, superiority of the CDSA over the conventional RAN architecture was critically discussed under several dimensions. These include: system capacity, EE, interference, resource and mobility management. In the capacity dimension, several capacity expansion mechanisms were classified and their pros/cons were discussed under both CDSA and conventional RAN configurations. In the EE dimension, several energy saving techniques were reviewed and the always-on service approach was found as the main cause of load-insensitive EC profiles. In contrast, the relaxed coverage constraints in the CDSA resulted in an on-demand always-available system that achieves a breakthrough in energy saving. The centralised CP enabled by the CDSA was found to be an important factor in designing efficient interference management techniques and network-driven resource assignment strategies. Such mechanisms are of great importance in dense SC deployment scenarios. Regarding mobility management, the dual connectivity feature of the CDSA enabled robust mobility performance with relaxed constraints. Despite these advantages, the CDSA imposes new research and technical challenges. In this direction, context information and the concepts of SON and SDN can play a key role. Based on this survey, two important areas, namely signalling overhead and mobility management, were identified as potential aspects that can be improved under CDSA networks, but they lack of novel techniques tailored for the CDSA.

Subsequently, Chapter 3 focused on the in-band signalling related to channel estimation. The inefficiency of the conventional architecture in terms of in-band and PL signalling was discussed and linked to the worst-case design approach. The latter can be traced to the fact that all users of the conventional architecture are connected to the same BS irrespective of their activity state. In the CDSA, however, only the active UE are associated with the DBS which lends the UE-DBS link to flexible and adaptive allocation techniques. Therefore, an adaptive DBS pilot signalling scheme was proposed in Chapter 3 to minimise the CDSA in-band signalling used for channel estimation in
the DBS layer. The proposed scheme considers the DL of a multi-carrier air interface and it takes into account both FD and TD variations. It depends on estimating the actual channel correlation function under realistic conditions rather than considering the worst-case CCs. To remove the noise effects from the estimated correlation function, a noise compensation unit based on a polynomial fitting approach was proposed. A theoretical comparison between the proposed adaptive scheme and the conventional worst-case design indicated promising gains of more than 90% reduction in the pilot overhead in LA scenarios. In addition, link level simulations were performed to assess gains and performance penalty (if any) of the proposed scheme over the LTE CRS pattern. The results showed that the proposed DBS adaptive pilot signalling scheme reduces the pilot overhead by 74%–78% w.r.t. the LTE CRS pattern without (with a marginal) performance penalty in low (high) SNR regions.

In Chapter 4, the out-of-band signalling related to mobility management was investigated. The DBS HO signalling was classified into three main components: air interface, RAN and CN signalling. The first two components were tackled in Chapter 4 by proposing predictive DBS HO schemes with 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 MRs 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. In this context, two predictive DBS HO signalling schemes were proposed: a history-based predictive DBS HO signalling and a measurement-based context-aided predictive DBS HO signalling. The former predicts future HO events and performs the RAN signalling in advance based on the UE HO history. This scheme includes an online learning and prediction process that consists of computationally simple arithmetic operations with minimal storage requirements. In addition, it can be used as an initial step in solving the DP holes problem in energy efficient CDSA networks with dynamic on/off DBS operation. On the other hand, the measurement-based context-aided predictive DBS HO signalling scheme enables timely advance RAN signalling by predicting both future HO events and the
expected HO time. It is operated at the UE side and combines SS/SQ measurements to physical proximity and UE context information in terms of location, direction and HO history. To improve the prediction performance whilst minimising the processing load, a UE-specific triggering condition was proposed and modelled as a function of the UE speed, the HO hysteresis and the network parameters. Both predictive schemes consider a dual mode operation by allowing the UE to revert back to the conventional HO signalling procedure. The proposed predictive schemes were implemented in a system level simulator, and the results showed 34% reduction in the HO latency as compared with the conventional non-predictive HO approaches.

Chapter 5 focused on the CN signalling part of the HO process by exploiting the dual connectivity feature of the CDSA. The fact that the UE is always anchored to the CBS allows the latter to act as an RRC anchor point for the UE and as a DP anchor point for the DBSs. This approach resulted in a CN-transparent HO signalling scheme for intra-CBS HOs because the CP and the CN-CBS DP path remain unchanged as long as the UE mobility is within the same umbrella. An analytical framework based on stochastic geometry was developed to model the CN signalling load as a function of network deployment parameters, UE contextual information and session characteristics. The CN signalling load and probability were derived for generic distributions, and closed-form expressions were obtained for the case where the cell residence time and the session duration are exponentially distributed. Moreover, the additional DP latency due to the CBS hop was modelled and an upper bound for the DBS density under latency constraints was derived. Numerical and simulation results indicated that the proposed CN-transparent HO signalling scheme is more beneficial in dense deployment scenarios where the gain reaches 90% reduction in the CN signalling load w.r.t. the conventional CN-visible HO signalling approaches. In addition, the CN-transparent HO signalling scheme was integrated with the predictive HO models of Chapter 4. System level simulation results showed that the integrated scheme reduces the HO signalling latency by 60% as compared with the conventional non-predictive and CN-visible HO techniques.
6.2 Future Work

In this section, future research directions and approaches to broaden the scope of this thesis and to extend this work are discussed.

6.2.1 Predictive Mobility Management with DBS Sleep Modes

In energy efficient CDSA networks with on/off DBS operation, the HO decision is not a trivial process. This can be traced to the CDSA system model and the functionality mapping in Table 2.1 where the best serving DBS (for the initial association) and the best candidate DBS (for the HO decision) may not be discoverable by the UE. Expressed differently, MRs of some DBSs may not be available when they are switched off for energy saving. This in turn constrains either resource selection and HO prediction strategies or sleep mode polices. In the former, the UE may not be able to measure signals of the inactive DBSs, thus the serving node selection decision will exclude these DBSs from the candidate set. Subsequently, the measurement-based context-aided predictive DBS HO signalling scheme in Section 4.4 will exclude the inactive DBSs from the prediction process. The history-based predictive DBS HO signalling scheme in Section 4.3 could offer a solution to this problem because it does not depend on instantaneous MRs. However, the optimal DBS on/off switching time remains an issue in this scheme.

This suggests sub-optimal sleep modes where the DBSs have to be switched on periodically to send pilot signals for measurement purposes [74], [75]. However, such an operation limits the period in which the DBS components can be turned off. This calls for the design of indirect measurement and prediction techniques to avoid the periodic transmission of pilot signals, thus overcoming limitations of the traditional sleep mechanisms whilst improving the prediction outcome. A future work in this direction is to integrate the predictive HO signalling schemes in Chapter 4 with efficient (where possible, self organising) location/fingerprint correlation and interpolation techniques. The main objective of the latter is to estimate SS and/or SQ of the inactive DBSs based on historical MRs collected from several UE at nearby locations. The estimation outcome can then be used as an input to the prediction process. Such an approach can
be viewed as a collaborative predictive HO signalling scheme, where each UE utilises measurements of other UE to predict future HO events whilst maintaining the DBSs in a sleep mode (as long as they are not needed for data transmission).

### 6.2.2 Predictive Mobility Management for CDSA-based Multi-RAT

The new 5G use cases, diverse application requirements and network deployment scenarios have lead to a widely accepted assumption that a single radio access technology (RAT) paradigm may not be suitable in the 5G era [28]. According to recent discussions in standardisation, industrials and academic forums, the 5G system will be based on a highly integrative approach connecting any new air interface and RAN to legacy cellular technologies (such as the LTE) and other wireless standards (such as the WiFi) [1], [159]. Such an approach comes with a range of benefits: it increases the system capacity, the users will have more access options and different services can be provided by different RATs. Nonetheless, the multi-RAT approach raises new research and technical challenges in several fields.

Considering the CDSA concept, the DP can be supported by multiple RATs under the umbrella of a common CP [160], [161]. The latter performs centralised traffic offloading/steering between different RATs under the CDSA theme of network-driven resource management. In this direction, the conventional reactive mobility management approach may not be suitable due to the expected increase in HO signalling load, interruption time, reporting overhead and measurement gaps. Even for static users, the multi-RAT traffic offloading and steering may create inter-RAT mobility events which are more complex than the conventional intra-RAT procedures. A more convenient design approach is to consider proactive mobility management techniques that exploit predictive and contextual-aware mechanisms rather than depending solely on instantaneous signal, channel and interference conditions. In this context, the proposed predictive HO schemes in Chapter 4 can be extended to enable efficient and optimised traffic steering and offloading mechanisms in multi-RAT scenarios under CDSA configuration.
6.2.3 Coordinated Adaptive Pilot Signalling Pattern for Multiple Antennas and Neighbouring DBSs

The adaptive DBS pilot signalling pattern proposed in Chapter 3 provides a significant saving in pilot overhead as compared with the conventional worst-case design approach. This scheme can be extended for other scenarios such as DBSs equipped with multiple antennas and neighbouring DBSs. In these scenarios, the conventional approach suffers from severe limitations. In the LTE for example, when an antenna port transmits a pilot symbol, other antenna ports mute their transmission completely during the resource element occupied by the pilot symbol of the neighbouring antenna to avoid pilot contamination [120]. Notice that the muting is performed only at the resource element level not at the frame level. Thus the overall pilot overhead increases by a factor of roughly the number of antennas. It would be interesting to investigate the effect of the adaptive DBS pilot signalling scheme on the pilot contamination problem, and whether the adaptive pattern requires or does not require resource element muting. This study might suggest a coordinated design approach, where neighbouring DBSs or antenna ports of the same DBS jointly decide the optimal adaptive pattern that reduces the pilot overhead without excessive resource element muting. In this direction, the Cloud-RAN [50] and the RME [90] concepts can be adopted to define the coordination rules and the pilot allocation constraints.

6.2.4 Self-organised Network-driven User Association in CDSA

The survey of the CDSA potential benefits presented in Chapter 2 draws a key conclusion on the importance of network-driven resource assignment strategies. In contrast to the conventional architecture, network access and service provisioning are supported by different nodes in the CDSA, namely the CBS and the DBS respectively. This allows the former to select the best serving DBS with a wide view of network status and parameters such as EC, congestion and interference, in addition to the knowledge of the UE mobility pattern and QoS requirements. Although this approach could offer an efficient resource management, optimising the association decision is not trivial. Identifying and prioritising the optimisation objectives is a challenging task due to the
trade-offs involved and the dynamic nature of operational networks. For instance, a mobility and HO signalling conscious user association strategy could select the DBS with the highest probability the UE will not leave it quickly, irrespective of network status. On the other hand, an energy efficient user association strategy could lead to a load imbalance across the DBSs. Similarly, a resource assignment scheme with a primary target of maximising the data rate could degrade the EE.

These trade-offs suggest an adaptive DBS-UE association strategy with a joint optimisation in signalling, mobility, energy, load balance and capacity dimensions. The term adaptive means that the optimisation objectives are dynamically prioritised according to the network status rather than adopting a static order. Several constraints need to be considered such as QoS requirements and, whenever relevant, the UE mobility pattern (which can be obtained from the prediction schemes in Chapter 4) should be incorporated in the constraints.
Bibliography


of IEEE International Conference on Communications (ICC), June 2015, pp. 2233–2238.


small cells with D2D links,” in *Proc. of IEEE International Conference on Communications (ICC)*, June 2013, pp. 1155–1160.


[114] 3GPP, Huawei, HiSilicon, “Feasible scenarios and benefits of dual connectivity in small cell deployment,” Written Contribution, 3GPP TSG-RAN WG2 Meeting


