CROSSFIRE
Uncoordinated network strategies for enhanced interference, mobility, radio resource, and energy saving management in LTE-Advanced networks

FP7 Contract Number: 317126

Deliverable D3.2: Component level mechanisms and algorithms for SONs and cognition in LTE-A

WP3

<table>
<thead>
<tr>
<th>Contractual Date of Delivery:</th>
<th>February 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Date of Delivery:</td>
<td>February 2015</td>
</tr>
<tr>
<td>Responsible Beneficiary:</td>
<td>IQU</td>
</tr>
<tr>
<td>Contributing Beneficiaries:</td>
<td>KCL, UOC, NEC, UoA</td>
</tr>
<tr>
<td>Security:</td>
<td>Public</td>
</tr>
<tr>
<td>Nature</td>
<td>Report</td>
</tr>
<tr>
<td>Version:</td>
<td>V1</td>
</tr>
</tbody>
</table>
Executive Summary

This document shows the activities carried out within the consortium about the algorithms for SON and cognition in LTE-A communication system. Each ESR tackled a different topic within the field and provided excellent contributions that paved their way into technical publications. Each chapter will present the achievements within each selected topic, where a review of the state of the art in first introduced in order to realize the current status on the science in the research arena, and to show the advantages of our proposals in comparison to the SoA.

In relation to D2D communication, we provide an alternative way to examine the existence of only cell-edge D2D links that cross due to their need to communicate different serving regions in cellular networks, and that helped us to understand and develop an efficient optimization framework to achieve advanced BS association policies as well as signalling and load balancing improvement. A set of numerical investigations is presented which shed light on the efficiency of our schemes to provide reliable BS association and load balancing performance, when active communication is taking place between any constructed pair of D2D nodes within the network.

Another very important topic within SON is the handovers optimization, where our work has been focused on the analysis and design of efficient Radio Resource Management (RRM) algorithms in the framework of SONs. In particular, we provided solutions aiming at improving the performance of the handover (HO) process in challenging heterogeneous networks. We first study the impact of the distance between the center of the macro and small cell on the HO performance, and we optimize the selection of the TTT value to improve the system behavior. We prove the dependency of the HO performance on the inter-site distance in a two-tier network. And we derive closed-form expressions for the different HO performance metrics as a function of inter-site distance and speed of UEs.

Dealing with the traffic offloading in small-cells as a Service (SCaaS) approach, we tackle this issue by considering financial aspects, where the distribution of the existing resources, i.e., the available spectrum and the cost of using the SCO infrastructure, are the response to two main objectives: the maximization of the throughput (the capacity objective), and the optimization of the profits (the economic objectives of both the MSPs and the SCO). We present the interaction of these competing objectives as the optimization process through an auction. We further address the interaction of the capacity needs and the economic constraints of the stakeholders, as we gain insight into techno-economic implications of the SCaaS paradigm.

An interesting topic is the cognition for Elastic Resource Sharing in Integrated FDD/TDD LTE-A HetNets, where we consider the amount of resources shared in the FDD system and the requirements of the users associated with the TDD picocells. Focusing on TDD systems, the idea of flexible spectrum sharing is investigated and we consider elastic resource sharing in a multi-tenant multi-operator LTE-A HetNet environment. We present the proposed re-configuration algorithm for the TDD frames, which enables efficient resource sharing between TDD system of pico eNBs and corresponding FDD system of macro eNBs.

The last subject tackled by the ESRs is the standardization and resource sharing optimization among D2D communications, where we focus on inband, overlay and unicast D2D links that are user-originated and controlled by the operator. We study the problem of D2D resource allocation in LTE-A networks, targeting at minimizing the amount of spectrum required for serving a specific number of overlaying D2D requests. The proposed scheme exploits the Ant Colony Optimization (ACO) theory and a graph representation of D2D mutual interferences, as a means to guarantee multiple concurrent D2D transmissions with specific target outage probability.
Table of Contents

List of Figures .......................................................................................................................................................... 5
List of Tables ............................................................................................................................................................ 8
Document History ...................................................................................................................................................... 9
List of Abbreviations and Acronyms ..................................................................................................................... 10

1. Algorithms for SONs and cognition in LTE-A for device to device communications ........................................ 13
   1.1 State-of-Art analysis ........................................................................................................................................... 13
       1.1.1 Device to Device Resource Allocation algorithms ................................................................................. 14
   1.2 Robust Randomized Resource Allocation ................................................................................................. 15
       1.2.1 Scenarios, methodology and performance analysis .............................................................................. 15
   1.3 Optimal Device-to-Device Cell Association and Load Balancing ............................................................. 21
       1.3.1 Scenarios, methodology and performance analysis .............................................................................. 22
   1.4 Future research planning .............................................................................................................................. 34

2. Algorithms for SONs and cognition in LTE-A for handovers optimization .................................................. 35
   2.1 Introduction ....................................................................................................................................................... 35
   2.2 Handover procedures ....................................................................................................................................... 36
   2.3 Handover Performance in LTE-A HetNets through Inter-Site Distance Differentiation ............................. 38
       2.3.1 Scenario and Methodology ....................................................................................................................... 38
       2.3.2 Numerical Results ....................................................................................................................................... 41
       2.3.3 Performance results ................................................................................................................................... 43
   2.4 The Impact of Inter-Site Distance and Time-to-Trigger on Handover Performance in LTE-A HetNets 46
       2.4.1 Scenario and Methodology ....................................................................................................................... 46
       2.4.2 Mathematical Analysis ............................................................................................................................ 48
       2.4.3 Numerical Results ....................................................................................................................................... 52
       2.4.4 Performance results ................................................................................................................................... 54
   2.5 Work plan and Future results ......................................................................................................................... 56
       2.5.1 Self-Organizing Users ............................................................................................................................... 56
       2.5.2 Future objectives ......................................................................................................................................... 57

3. Algorithms for SONs and cognition for Traffic Offloading with small-cells................................................. 59
   3.1 State State of the Art on traffic offloading with small-cells ........................................................................... 59
       3.1.1 Financial aspects .......................................................................................................................................... 59
   3.2 A novel learning mechanism for traffic offloading with Small Cell as a Service ........................................ 64
       3.2.1 Scenarios, methodology and performance analysis .............................................................................. 64
   3.3 Network and financial aspects within the SCaaS approach .......................................................................... 74
       3.3.1 Scenarios, methodology and performance analysis .............................................................................. 74
   3.4 Future research planning .............................................................................................................................. 83

4. Algorithms for SONs and cognition for decentralized network management methods, policies and algorithms ................................................................................................................................. 84
   4.1 SON Introduction ............................................................................................................................................... 84
   4.2 SON Standardization Status: ........................................................................................................................ 85
       4.2.1 Multi-Vendor SON: .................................................................................................................................... 90
       4.2.2 SON Architecture Implementation in Multi-Vendor environment .......................................................... 91
       4.3.1 Introduction ............................................................................................................................................... 93
       4.3.2 Related Work ........................................................................................................................................... 94
4.3.3 Elastic Resource Management in LTE-A HetNets

5. Towards the standardization and resource sharing optimization among D2D communications

5.1 Resource sharing over D2D: introduction and state of the art

5.2 Ant Colony Optimization for Resource Sharing among D2D Communications

5.2.1 System Model

5.2.2 Ant Colony Optimization

5.2.3 Evaluation Results

5.2.4 Conclusions

5.3 Enabling D2D communications in LTE networks

5.3.1 System Model

5.3.2 LTE Functionality for D2D Communications

5.3.3 Evaluation Results

5.3.4 Conclusions and Potential Applications

5.4 Future work

6. Conclusion

References

List of Figures

Figure 1: Cisco forecasts 24.3 exabytes of global mobile data traffic per month by 2019.

Figure 2: Topology / system model of D2D communications in LTE-A based cellular network. FRF = 1 is applied for inner users, whereas FRF = 3 is applied for cell-edge UEs.
Figure 3: Comparison among existing works and proposed scheme: "1" stands for Random Allocation, "2" for strict-FFR based Allocation, "3" for the differentiated FFR Allocation scheme in [11] and "4" for proposed scheme. ................................................................. 19

Figure 4: Aggregate throughput estimation related to the number of deployed D2D links. ................................................................. 20

Figure 5: Device-to-Device communication links in a multi-cell environment (with $h = 2d$). A number of links might cross the boundaries between neighbouring cells, hence link nodes should be connected to different BSs. ................................................................. 22

Figure 6: Buffon's Needle survey and Monte Carlo simulations comparison. This figure signifies the probability of having D2D link crossing within a pair of neighbouring cells. Convergence achieved for high number of iterations. ................................................................. 23

Figure 7: Different scenarios when nodes of a D2D are located on cell edge and communication with two BSs is possible. ................................................................. 25

Figure 8: Normalized cost comparison for cost minimization IP solver and Cost-based Heuristic (CbH) algorithm for high congestion levels. ................................................................. 30

Figure 9: Normalized utilization balancing comparison for optimal allocation of RBs. ................................................................. 31

Figure 10: Simulation instances of Load Balancing (LB) optimization solver, Load-based (LbH) and Cost-based heuristic (CbH) algorithmic solutions. ................................................................. 32

Figure 11: Mean standard deviation estimations for Load Balancing (LB-IP) solver, Load-based (LbH) and Cost-based heuristic (CbH). ................................................................. 33

Figure 12: Mean standard deviation estimations based on different BSs' utilization cases (UC) for Load Balancing (LB-IP) solver, LbH and CbH. ................................................................. 34

Figure 13: Global Mobile Devices and Connections Growth [1]................................................................. 35

Figure 14: Exabytes per Month of Mobile Data Traffic by 2019 [1] ................................................................. 36

Figure 15: Macro-Small cell HO scenario ................................................................. 39

Figure 16: No HO probability, speed = 60 km/h ................................................................. 43

Figure 17: inbound HO probability, speed = 60 km/h ................................................................. 44

Figure 18: RLF probability, speed = 60 km/h ................................................................. 45

Figure 19: PP probability, speed = 60 km/h ................................................................. 45

Figure 20: HOF probability, speed = 60 km/h ................................................................. 46

Figure 21: Small cell area and considered angles ................................................................. 48

Figure 22: HO Probability ($P_{\text{HO}}$) ................................................................. 54

Figure 23: HOF Probability ($P_{\text{HOF}}$) ................................................................. 55

Figure 24: RLF Probability ($P_{\text{RLF}}$) ................................................................. 56

Figure 25: PP Probability ($P_{\text{PP}}$) ................................................................. 56

Figure 26: $\textbf{MSP1}$ Estimation performance ................................................................. 74

Figure 27: $\textbf{MSP1}$ Throughput and Unserved traffic ................................................................. 74

Figure 28: $\textbf{MSP1}$ Profit and Bid ................................................................. 74

Figure 29: System Model and Spectrum Allocation for the two channel deployments ................................................................. 75

Figure 30: Traffic served by the SC cluster versus the total offered load for $\textbf{MSP1}$ ................................................................. 79

Figure 31: $\textbf{MSP1}$’s throughput versus the total offered load ................................................................. 79

Figure 32: $\textbf{MSP1}$’s profit versus the total offered load ................................................................. 80

Figure 33: $\textbf{MSP1}$ profit-throughput ratio versus $\textbf{MSP2}$’s total offered load for $L1 = 550\,\text{Mbps}$ ................................................................. 80

Figure 34: $\textbf{MSP1}$ throughput for no competition ($L2 = 0$) ................................................................. 81

Figure 35: $\textbf{MSP1}$ throughput for saturated competition ................................................................. 81

Figure 36: Throughput comparison for no competition ($L2 = 0$) ................................................................. 82

Figure 37: Throughput comparison for saturated competition ................................................................. 82

Figure 38: System Model ................................................................. 83

Figure 39: 3GPP Management Reference Model ................................................................. 91
Figure 40: Distributed SON ................................................................. 92
Figure 41: Vendor A Centralized and Vendor B Distributed SON ................ 93
Figure 42: SDN-based network resource management architecture .................. 97
Figure 43: Logic Signaling Flow Mechanism ............................................. 98
Figure 44: Application Layer Delay CDF of the FDD Macro eNB System ............ 101
Figure 45: Application Layer Delay CDF of the TDD Pico eNB System .......... 102
Figure 46: D2D classification types ......................................................... 103
Figure 47: a) A snapshot of D2D transmitters’ locations, b) The corresponding total interference inside a cell with multiple concurrent D2D transmissions. .......................................................... 108
Figure 48: System model example, $D = 9$, $S = 3$, $L_1 = 4$, $L_2 = 3$, $L_3 = 2$. ................................................................. 109
Figure 49: Interference graph of the D2D pairs. ............................................. 110
Figure 50: Gradual convergence of ACO algorithm to a close-to-optimal solution. ................................................................. 114
Figure 51: Achieved spectrum reuse by the ACO approach compared to optimal. ................................................................. 115
Figure 52: Outage probabilities for various subset lengths ................................ 116
Figure 53: D2D Model ........................................................................... 118
Figure 54: D2D enhancements in standard LTE signalling .............................. 119
Figure 55: Enhanced MAC PDU .......................................................... 120
Figure 56: SINR at eNB for different number and locations of D2D transmitters. ................................................................. 122
Figure 57: Spatial spectrum reuse ratio in whole cell for different UE locations. ................................................................. 123
### List of Tables

Table 1: Simulation parameters and typical values ................................................................. 18  
Table 2: Sensitivity analysis related to location uncertainty.................................................. 21  
Table 3: Simulation parameters ............................................................................................ 29  
Table 4 Simulation Parameters ............................................................................................... 41  
Table 5 Simulation Parameters .............................................................................................. 53  
Table 6: MSP-SCO Network and Financial parameters ......................................................... 73  
Table 7: Learning Mechanism parameters ............................................................................. 73  
Table 8: MSP-SCO Network Financial Parameters ............................................................... 78  
Table 9: Average Spectral Efficiency for different small cell deployments ............................ 78  
Table 10: System level simulation parameters ...................................................................... 100  
Table 11: Basic simulation parameters. .................................................................................. 113  
Table 12: Basic simulation parameters. .................................................................................. 121
## Document History

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Status</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.2.2015</td>
<td>V0.1</td>
<td>Draft</td>
<td>First ToC version</td>
</tr>
<tr>
<td>23.4.2015</td>
<td>V0.2</td>
<td>Draft</td>
<td>Full version</td>
</tr>
</tbody>
</table>
### List of Abbreviations and Acronyms

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>Third (3rd) Generation Partnership Project</td>
</tr>
<tr>
<td>ACO</td>
<td>Ant Colony Optimization</td>
</tr>
<tr>
<td>ACP</td>
<td>ACcess Permission</td>
</tr>
<tr>
<td>AP</td>
<td>Access Point</td>
</tr>
<tr>
<td>APO</td>
<td>Access Point Owner</td>
</tr>
<tr>
<td>BN</td>
<td>Buffon Needle</td>
</tr>
<tr>
<td>BS</td>
<td>Base Station</td>
</tr>
<tr>
<td>BSR</td>
<td>Buffer Status Report</td>
</tr>
<tr>
<td>C2M</td>
<td>Cellular-to-mesh</td>
</tr>
<tr>
<td>CAPEX</td>
<td>Capital Expenditure</td>
</tr>
<tr>
<td>CBH</td>
<td>Cost based Heuristic</td>
</tr>
<tr>
<td>CCCH</td>
<td>Common Control Channel</td>
</tr>
<tr>
<td>CI</td>
<td>Complete Information</td>
</tr>
<tr>
<td>CR</td>
<td>Cognitive Radio</td>
</tr>
<tr>
<td>C-RNTI</td>
<td>Cell-Radio Network Temporary Identifier</td>
</tr>
<tr>
<td>CSI</td>
<td>Channel State Information</td>
</tr>
<tr>
<td>CU</td>
<td>Cellular User</td>
</tr>
<tr>
<td>D2D</td>
<td>Device-to-Device</td>
</tr>
<tr>
<td>DL</td>
<td>Downlink</td>
</tr>
<tr>
<td>DOFF</td>
<td>Data OFFloading game</td>
</tr>
<tr>
<td>DSR</td>
<td>Dynamic Source Routing</td>
</tr>
<tr>
<td>DU/De</td>
<td>UL/DL Decoupling</td>
</tr>
<tr>
<td>eNB</td>
<td>evolved NodeB</td>
</tr>
<tr>
<td>EPS</td>
<td>Evolved Packet System</td>
</tr>
<tr>
<td>eUE</td>
<td>enhanced UE</td>
</tr>
<tr>
<td>E-UTRAN</td>
<td>Evolved – UMTS Terrestrial Radio Access Network</td>
</tr>
<tr>
<td>FDD</td>
<td>Frequency Division Duplex</td>
</tr>
<tr>
<td>FFR</td>
<td>Fractional Frequency Reuse</td>
</tr>
<tr>
<td>FO</td>
<td>Fixed-line Operator</td>
</tr>
<tr>
<td>FRF</td>
<td>Frequency Reuse Factor</td>
</tr>
<tr>
<td>FSP</td>
<td>Femtocell Service Provider</td>
</tr>
<tr>
<td>FSSG</td>
<td>Femtocell Service Selection Game</td>
</tr>
<tr>
<td>H</td>
<td>Hysteresis Margin</td>
</tr>
<tr>
<td>HetNets</td>
<td>Heterogeneous Networks</td>
</tr>
<tr>
<td>HO</td>
<td>Handover</td>
</tr>
<tr>
<td>HOF</td>
<td>Handover Failure</td>
</tr>
<tr>
<td>HTC</td>
<td>Human Type Communications</td>
</tr>
<tr>
<td>HW</td>
<td>Holt-Winters</td>
</tr>
<tr>
<td>ILI</td>
<td>Interference Level Indicator</td>
</tr>
<tr>
<td>ILP</td>
<td>Integer Linear Programming</td>
</tr>
<tr>
<td>IMSI</td>
<td>International Mobile Subscriber Identity</td>
</tr>
<tr>
<td>IMT-A</td>
<td>International Mobile Telecommunications-Advanced</td>
</tr>
<tr>
<td>IP</td>
<td>Integer Programming</td>
</tr>
<tr>
<td>ITU-R</td>
<td>International Telecommunication Union – Radio communication sector</td>
</tr>
<tr>
<td>LB</td>
<td>Load Balancing</td>
</tr>
<tr>
<td>LbH</td>
<td>Load based Heuristic</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>LCID</td>
<td>Logical Channel ID</td>
</tr>
<tr>
<td>LIPA</td>
<td>Local IP Access</td>
</tr>
<tr>
<td>LTE-A</td>
<td>Long Term Evolution-Advanced</td>
</tr>
<tr>
<td>MC</td>
<td>Marginal Cost</td>
</tr>
<tr>
<td>MINLP</td>
<td>Mixed Integer Non Linear Programming</td>
</tr>
<tr>
<td>MIP</td>
<td>Mixed Integer Programming</td>
</tr>
<tr>
<td>MM</td>
<td>Mobility Management</td>
</tr>
<tr>
<td>MME</td>
<td>Mobility Management Entity</td>
</tr>
<tr>
<td>MO</td>
<td>Mobile Operator</td>
</tr>
<tr>
<td>MSIN</td>
<td>Mobile Subscriber Identification Number</td>
</tr>
<tr>
<td>MTC</td>
<td>Machine Type Communications</td>
</tr>
<tr>
<td>MTS</td>
<td>Minimum Time of Stay</td>
</tr>
<tr>
<td>NTP</td>
<td>Network Termination Point</td>
</tr>
<tr>
<td>OPEX</td>
<td>Operational Expenditure</td>
</tr>
<tr>
<td>OSM</td>
<td>Objective Sum Method</td>
</tr>
<tr>
<td>PC</td>
<td>Power Control</td>
</tr>
<tr>
<td>PCC</td>
<td>Policy and Charging and Control</td>
</tr>
<tr>
<td>PDCCH</td>
<td>Physical DL Control Channel</td>
</tr>
<tr>
<td>PDN</td>
<td>Packet Data Network</td>
</tr>
<tr>
<td>PDU</td>
<td>Protocol Data Unit</td>
</tr>
<tr>
<td>PO</td>
<td>Primary spectrum Owner</td>
</tr>
<tr>
<td>PP</td>
<td>Ping-Pong</td>
</tr>
<tr>
<td>PRACH</td>
<td>Physical Random Access Channel</td>
</tr>
<tr>
<td>ProSe</td>
<td>Proximity Services</td>
</tr>
<tr>
<td>PSCF</td>
<td>Proximity Service Control Function</td>
</tr>
<tr>
<td>PSM</td>
<td>ProSe Management</td>
</tr>
<tr>
<td>PU</td>
<td>Primary User</td>
</tr>
<tr>
<td>PUCCH</td>
<td>Physical UL Control Channel</td>
</tr>
<tr>
<td>PUSCH</td>
<td>Physical UL Shared Channel</td>
</tr>
<tr>
<td>QoE</td>
<td>Quality of Experience</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RAN</td>
<td>Radio Access Network</td>
</tr>
<tr>
<td>RB</td>
<td>Resource Block</td>
</tr>
<tr>
<td>RLF</td>
<td>Radio Link Failure</td>
</tr>
<tr>
<td>ROI</td>
<td>Region of Interest</td>
</tr>
<tr>
<td>RRC</td>
<td>Radio Resource Control</td>
</tr>
<tr>
<td>RRM</td>
<td>Radio Resource Management</td>
</tr>
<tr>
<td>RSRP</td>
<td>Reference Signal Received Power</td>
</tr>
<tr>
<td>RSS</td>
<td>Received Signal Strength</td>
</tr>
<tr>
<td>Rx</td>
<td>Receiver</td>
</tr>
<tr>
<td>SBS</td>
<td>Small Base Stations</td>
</tr>
<tr>
<td>SCaaS</td>
<td>Small Cell as a Service</td>
</tr>
<tr>
<td>SCO</td>
<td>Small Cell Operator</td>
</tr>
<tr>
<td>SIC</td>
<td>successive interference cancelation</td>
</tr>
<tr>
<td>SINR</td>
<td>Signal to Interference plus Noise Ratio</td>
</tr>
<tr>
<td>SIPTO</td>
<td>Selected IP Traffic Offload</td>
</tr>
<tr>
<td>SLA</td>
<td>Service Level Agreement</td>
</tr>
<tr>
<td>SON</td>
<td>Self-Organized Networks</td>
</tr>
<tr>
<td>SR</td>
<td>Scheduling Request</td>
</tr>
<tr>
<td>SSO</td>
<td>Small Scale Operator</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>SSP</td>
<td>Small cell Service Provider</td>
</tr>
<tr>
<td>S-TMSI</td>
<td>Subscriber-Temporary Mobile Subscriber Identity</td>
</tr>
<tr>
<td>SU</td>
<td>Secondary User</td>
</tr>
<tr>
<td>TDD</td>
<td>Time Division Duplex</td>
</tr>
<tr>
<td>ToS</td>
<td>Time of Stay</td>
</tr>
<tr>
<td>TTI</td>
<td>Transmission Time Interval</td>
</tr>
<tr>
<td>Tx</td>
<td>Transmitter</td>
</tr>
<tr>
<td>UE</td>
<td>User Equipment</td>
</tr>
<tr>
<td>UL</td>
<td>Uplink</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telecommunications System</td>
</tr>
<tr>
<td>VCG</td>
<td>Vickrey-Clark-Groves</td>
</tr>
<tr>
<td>WSP</td>
<td>Wireless Service Provider</td>
</tr>
</tbody>
</table>
1. Algorithms for SONs and cognition in LTE-A for device to device communications

1.1 State-of-Art analysis

The unquestionable need to encounter the rapidly increasing demand for wireless access to the existing and emerging mobile networks is the main concern for network operators. It is widely noticed that the average user owns at least one mobile device with enriched set of capabilities such as smartphones or tablets, while at the same time each one of these devices' usage grew in a percentage of 45% on average during 2014 according to a survey conducted by Cisco [1]. The figure shown below highlights the corresponding forecasts over the mobile data traffic growth until 2019.

![Cisco forecasts 24.3 exabytes of global mobile data traffic per month by 2019.](image)

This issue initiated the research community's interest to investigate for novel networking solutions in order to satisfy or further improve the requirements set in the IMT-Advanced specifications. Currently, 3GPP LTE-A system is the one to be considered as the commercially introduced 4G technology, including all technological components to support heterogeneity and enhance the system capabilities and performance. Additionally, future 5G networks are expected to address the aforementioned user inundation by becoming more dynamic and ultra-dense and at the same time ameliorate the already existent interference mitigation as well as spectrum efficiency techniques [2]. Device-to-Device (D2D) communication contributes to this direction by bringing a set of benefits that arise mainly from the proximity gain that it offers [3], as well as from the reuse gain that the underlay property can provide [4]. In this emerging communication paradigm, two close-ranged UEs are eligible to connect directly and communicate with each other by utilizing either the cellular (i.e. inband) or the unlicensed spectrum (i.e. outband), unlike the traditional communication via the BS. The aim of introducing such a technique is to enable opportunities for new services, to provide novel means to offload cellular BSs and enhance network capacity by exploiting the advantages that localized, short-range wireless peer-to-peer communications can offer.

However, due to the possible existence of an increased number of users in a deployed network, the main concern on the use of inband underlaying D2D concept in a cellular network is the interference that may result from the aforementioned reuse from the D2D UEs of the already assigned cellular radio resources. Such reuse of wireless resources in interference limited networks needs to take place in a controlled manner to avoid critical overall performance degradation due to the aggregated interference [5].
1.1.1 Device to Device Resource Allocation algorithms

Resource management for D2D communications can be categorized depending on the nature of spectrum sharing to centralized, semi-distributed [6] or fully-distributed [7] allocation. It is well known that the problem of optimal reuse of resource blocks (RBs) can be transformed as special case of a graph coloring problem. This implies an inherent combinatorial nature that for general graphs entails an NP-hard problem, which informally speaking relates to problems with no known polynomial solution. Therefore, optimal solutions via a mathematical programming formulation cannot be deemed as suitable for real time implementations. Randomized algorithms including meta-heuristics have been proven to be a powerful toolbox to provide competitive solutions for general NP-hard problems. Inspired by these set of algorithms we are proposing a randomized scheme which shows demonstrable improvement over previous works (over 10% average gain in terms of throughput) in a low complexity manner.

In order to improve scalability in D2D communications and increase the spectral efficiency of the network, reliable resource allocation, in collaboration with interference-aware mechanisms is required. Interference is one of the most significant issues to be addressed in underlaying inband D2D communications due to the resource reuse between cellular and D2D users. Interference mitigation/cancellation methods have been proven to be very important towards this direction, while they enhance the system capacity without changing the network infrastructure [8][9]. In the latter, Xu et al. propose an interference cancellation technique by firstly allocating dedicated control channel for all D2D users. Cellular users then listen to the signalling of this common channel and measure its signal-to-interference noise ratio (SINR), the BS monitors the acquired information and decides the cease of scheduling cellular UEs on the RBs that are used by the related D2D UE, just in case the measured SINR doesn't satisfy a predefined threshold. Then, the BS broadcasts the allocated resources and the position of the cellular UEs in the dedicated channel and helps D2Ds avoid using the same RBs and therefore cause interference. Similarly enough, in [8], the authors differentiate the same case (i.e. measuring the signalling power of cellular users and broadcasting this information to the BS) and minimize the maximum received power at D2D pairs from cellular users, succeeding to increase the mean cell capacity in comparison with a conventional cellular operation.

The authors in [10] formulate the problem of spectrum sharing to D2D communications as a mixed integer non-linear programming (MINLP) that is hard to solve in real time. Their scheme is based on a greedy heuristic algorithm to mitigate the interference levels in the primary cellular network by exploiting channel gain information. In [11], Chae et al. are, to the best of our knowledge, the first to introduce the concept of a differentiated FFR in D2D communications integrated into a cellular network in order to mainly mitigate inter-cell interference caused by cell-edge users. In summary, their scheme aims to improve the performance for both cellular and D2D pairs by controlling interference that is developed between them. Unlike the conventional strict FFR method [12], the rationale in this scheme is to allocate different resources to D2D users compared to the cellular users that are located in the same region.

Considering only a single-cell scenario in [13], Yu et al. analyse the mode selection issue along with a power control technique and optimum resource allocation for the involved users that are subjected to spectral efficiency and energy restrictions. They define the sum-rate of the D2D and cellular communications by applying the Shannon capacity formula and with no specific assumptions on the background cellular networks. They firstly study a greedy sum-rate maximization problem where the two communication types compete services. Then, they set some priorities to the cellular users for obtaining the available radio resources with a guaranteed minimum transmission rate.

In [14], Phunchongharn et al. propose a joint RB scheduling and power control method to maximize the overall spectrum utilization by finding the minimum transmission length in terms of used timeslots for D2D pairs by guaranteeing at the same time their QoS minimum restrictions. In order to achieve this, they formulate their problem as a mixed integer programming (MIP) and make use of a scheme that is based on a column generation method to solve the resource allocation problem.
1.2 Robust Randomized Resource Allocation

1.2.1 Scenarios, methodology and performance analysis

In this section, we firstly introduce the scenario and the system model of our investigation and then proceed by outlining the proposed randomized radio resource allocation scheme.

1.2.1.1 System model and investigated scenarios

We consider a 7-cell scenario with center-deployed base stations and multiple cellular and D2D users, uniformly distributed in each cell. The D2D users are met in close-distanced pairs and satisfy their transmission needs by reusing the cellular network's available spectrum resources. Without loss of generality, all the deployed network elements, i.e. base stations and terminals are equipped with omnidirectional antennas. In this case, we investigate the downlink case (DL) and assume that, in every cell, $C$ cellular users occupy $C$ of the total number of $N$ available orthogonal sub-channels and $D$ D2D pairs can be either allocated a dedicated resource that is unused or reuse a cellular resource. We further assume that all users satisfy their minimum SINR restrictive requirements.

![Figure 2: Topology / system model of D2D communications in LTE-A based cellular network. FRF = 1 is applied for inner users, whereas FRF = 3 is applied for cell-edge UEs.](image)

In Figure 2, we partly represent the downlink scenario with some indicative D2D pairs and cellular UEs distributed in the deployed topology. Complying with the assumptions of the authors in [11], different interference scenarios for a D2D pair can be developed, depending on whether it is located in the inner or the outer region of a cell. For example, in cell 3, the outer region D2D receiver (Rx) can be interfered by the center-located BS that concurrently transmits to the inner-positioned CU by the time they use the same resource. Accordingly, a cellular UE's data reception can be harmed by the simultaneous transmission of the D2D pair that utilizes the same channel with it. The figure above depicts a part of this scenario, where the total bandwidth is divided to four orthogonal sub-channels $F1,F2,F3,F4$ that contain different available resources for the users located in each corresponding area, with the frequency reuse factor (FRF - rate at which the same frequency can be used in the network) set to one for inner users and three for the cell-edge users. Due to the aforementioned orthogonalized...
properties, proper resource assignment among CU and D2D UEs would be able to ease network operators to succeed an improved spectrum utilization and highlight the contribution of D2D in achieving increased rates for devices in close proximity which can be translated as accumulative gains in terms of aggregate throughput.

The path-loss models applied to the aforementioned scenario are defined in accordance with \cite{11} and are different for cellular and D2D UEs:

\[ PL_{CU} = 36.7 \log_{10}(d) + 40.9 + 26 \log_{10} \left( \frac{f_c}{5} \right) \]  \tag{1}

\[ PL_{D2D} = 40 \log_{10}(d) + 30 \log_{10}(f_c) + 79 \]  \tag{2}

where \( d \) is the distance (in meters) between the transmitting BS and a CU receiver for equation (1) and the distance between the D2D transmitter and his paired receiver for equation (2), respectively, whereas \( f_c \) is the carrier frequency expressed in GHz. Also, average rates are assumed, hence fast fading impairments are not taken into account.

Due to the concurrent transmissions of all deployed BSs towards their assigned CUs and from the D2D transmitters to the D2D receivers in the downlink period, severe interference cases may deteriorate the received rates of both user types. The transmitting power of each BS may be harmful for the receiver of a D2D pair that uses the same resource assigned to a cellular user. Additionally, if a second D2D pair reuses the same resource as aforementioned, it can also cause increased interference level at the receiving D2D UE. On the other hand, the receiving rates of cellular users can be influenced by the transmission power of a D2D transmitter that reuses the same sub-channel. The received SINR of a user \( i \) in \( l \)-th cell during the downlink period is given by:

\[ SINR_{i,l} = \frac{G_{i,l} P_{T_i}}{\sigma^2 + \sum_{z \in Z_{i,l}} (\sum_{q \in Q} G_{q,z} P_q)} \]  \tag{3}

where \( i \) stands for either \( c \) if the user is cellular or \( d \) if it corresponds to a D2D user, \( G_{i,l} \) is the channel gain for user \( i \) from the serving BS \( l \), \( P_{T_i} \) is the transmission power of either a BS or D2D transmitting user \( i \); in the denominator, \( \sigma^2 \) notes the noise power, \( Z_{i,l} \) is the set of cells that contain interfering nodes that use the same sub-band as user \( i \) of cell \( l \), \( Q \) is the set of interferers (including both BSs or D2D transmitters that cause interference) and \( P_q \) is the transmission power of interferer \( q \in Q \) and refers to either a BS's or D2D Tx's transmission power.

Each user's received SINR can be then mapped to a rate according to the Shannon capacity formula and can be expressed as follows:

\[ R_{i,l} = B \log_2 \left( 1 + SINR_{i,l} \right) \]  \tag{4}

where \( i \) corresponds to either \( c \) or \( d \), \( B \) is the sub-channel bandwidth and \( R_{i,l} \) equals to the estimated rate of user \( i \) in cell \( l \). Consequently, the system’s aggregate rate is given by:

\[ R_{tot} = \sum_{l=1}^{L} \sum_{i=1}^{U_l} R_{i,l} \]  \tag{5}

where \( U_l \) is the set of all deployed (i.e. both cellular and D2D) users in cell \( l \).
1.2.1.2 Methodology

Regarding the cellular users, the assumption is that they are firstly allocated $C$ orthogonal RBs in every cell. The available RB pool for users located in inner region ($N_{\text{inner}}$) is proportional to the interior-area radius and is twice the size of the pool corresponding to the cell-edge users ($N_{\text{outer}}$) [12]. Then, according to the location of the deployed D2D UEs, we randomly assign one RB for each D2D pair transmission needs, following the resource allocating rationale that is stated in [11], and then we subtract it from the related RB pool. In case that all RBs are occupied, a cellular resource can be occupied and reused by more than one D2D pair. This problem can be mathematically formulated according to [10] in order to maximize the sum-rate of CU and D2D UEs subjected to some performance satisfaction constraints and with the assumption of allocating only one RB to each user.

Because of the small number of available RBs in practice and looking for example in a specific sub-area of the different FFR cases, we apply a randomized algorithm which runs in a decentralized (or centralized) manner in all D2D transmitting nodes as follows: we assume that we have $N$ available RBs in the scope area, either it is for an inner or outer cell region. Then, each BS will randomly allocate RB to different D2D nodes up to a designated number of iterations. This is done for each D2D transmitter and the outcome is announced to the BS. Then, the base stations will select the best allocation out from all the ones calculated. The criterion of selecting the best out of this number of iterations allocation scheme is the aggregate throughput of both CU and D2D UEs in all cells, implying an effective in terms of system rate and interference controllability method. The algorithmic steps are given in Algorithm 1 that is presented below. Note that due to its nature, the proposed scheme falls within the category of "embarrassingly" parallel problems because iterations of the algorithm to explore the search space can be executed without requiring any communication between them [15]. In addition, the complexity of the proposed scheme is $O(K)$, i.e. linear increase with respect to the number of iterations $K$. 
Algorithm 1: Randomized RRA algorithm for D2D links.

1.2.1.3 Performance evaluation

We consider a LTE-A network scenario with 7 hexagonal cells, where cellular UEs communicate with a center-located BS and coexist with multiple D2D users in each cell area. In our base simulations, each cell serves 30 cellular UEs and 24 D2D pairs that are ranged in a minimum distance of 200 meters away from the BS. All involved users are being finally allocated with only one RB for their communication needs. The most important system level parameters used in the simulations and their corresponding values are summarized in Table 1.

Table 1: Simulation parameters and typical values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hexagonal cells</td>
<td>7</td>
</tr>
<tr>
<td>Number of cellular UEs per cell</td>
<td>30</td>
</tr>
<tr>
<td>BS transmission power</td>
<td>43 dBm</td>
</tr>
<tr>
<td>UE transmission power</td>
<td>15 dBm</td>
</tr>
<tr>
<td>Noise power density</td>
<td>-174 dBm/Hz</td>
</tr>
<tr>
<td>Incircle radius</td>
<td>800 m</td>
</tr>
<tr>
<td>Parameter</td>
<td>Value</td>
</tr>
<tr>
<td>----------------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Inter-BS distance</td>
<td>1600 m</td>
</tr>
<tr>
<td>D2D-pair distance range (uniformly distributed)</td>
<td>[20,55] m</td>
</tr>
<tr>
<td>Carrier frequency</td>
<td>2 GHz</td>
</tr>
<tr>
<td>System bandwidth</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Bandwidth of subchannel</td>
<td>180 kHz</td>
</tr>
<tr>
<td>Number of subchannels</td>
<td>50</td>
</tr>
</tbody>
</table>

The simulation results presented hereafter focus only on the case where D2Ds can reuse the DL resources of the cellular network. The decision of allocating specific resources to the deployed UEs according to FFR depends on their location and is characterized by their distance from the related BS. The comparisons and gains of this algorithmic implementation are presented in terms of aggregate throughput as shown in Eq. (5).

We evaluate the performance of the proposed randomized D2D RRA algorithm compared to three different allocation schemes. In the first scheme, resources are randomly allocated to all cellular and D2D users from the entire frequency band, whereas the second scheme follows the cellular resource allocation rationale for D2D users that can select an RB from the same frequency sub-band that CUs can choose. In the third scheme, a different resource allocation pool is available to D2D pairs (determined by their location in either inner or outer cell areas) to limit interference as described afore [11]. Our proposed method’s throughput is estimated via Monte Carlo simulations and its mean value is compared with the average, worst and best achievable system rates for the rest of the comparing schemes. As shown in Figure 3, the proposed randomized scheme manages to achieve a gain of 9.8% in terms of mean aggregate throughput (for 1000 iterations) compared to the baseline and over 12.5% of its minimum value, something that can be explained by the iterative round of checks for finding the best RB allocation combination for the deployed D2D UEs. Additionally, the main difference is observed in the rate performance of D2D users that is reasonably enhanced compared to the other techniques.

![Figure 3: Comparison among existing works and proposed scheme: "1" stands for Random Allocation, "2" for strict-FFR based Allocation, "3" for the differentiated FFR Allocation scheme in [11] and "4" for proposed scheme.](image)
Figure 4 illustrates the overall system throughput in relation to different numbers of D2D pairs per cell. We can see that throughput increases proportionally with the increase of D2D pairs' number, while, for the same data, random allocation method has a slightly increasing performance but rather abnormal behavior for increased number of D2D pairs, whereas the strict-FFR based allocation is characterized by a decreasing tendency in more dense scenarios due to the constrained resource reuse, especially for cell-edge D2D users. The benefits of the proposed randomized RRA algorithm are more distinctly observed in higher density scenarios where its maximum achievable gain in comparison with the other schemes (i.e. 30 D2D pairs per cell) is more than 91% compared to random allocation scheme, 60% compared to the strict-FFR implementation and 14.5% to the differentiated-FFR resource allocation scheme, respectively (for the case of 30 D2D pairs per cell).

![Figure 4: Aggregate throughput estimation related to the number of deployed D2D links.](image)

In addition, we performed a sensitivity analysis of the proposed scheme related to user location uncertainty and the results are graphically presented in Table 2. In this study, we firstly estimate the aggregate throughput when we have full knowledge of the location settings of all deployed terminals and then compare it with some cases where the nodes can be erroneously met in different location within a defined distance. By retaining the same RB assignment for all instances, the computed differentiation in terms of system throughput was met in a percentage of less than 1% for the investigated cases within the range of [1,5] meters. Also, unlike the throughput decrease that derives from the error in location accuracy, in some cases the aggregate throughput can be increased as the algorithm allocates the available resources without knowing the exact position coordinates of the involved nodes (i.e. within 4 and 5 meters the throughput is 0.55% and 0.18% improved, respectively). This table indicates the robustness of this algorithm to low uncertainty phenomena and its ability to sustain the network performance in terms of throughput.
Table 2: Sensitivity analysis related to location uncertainty.

<table>
<thead>
<tr>
<th>Location information error</th>
<th>Estimated aggregate throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accurate location</td>
<td>5.58*10^8</td>
</tr>
<tr>
<td>1 meter</td>
<td>5.51*10^8</td>
</tr>
<tr>
<td>2 meters</td>
<td>5.58*10^8</td>
</tr>
<tr>
<td>3 meters</td>
<td>5.56*10^8</td>
</tr>
<tr>
<td>4 meters</td>
<td>5.61*10^8</td>
</tr>
<tr>
<td>5 meters</td>
<td>5.59*10^8</td>
</tr>
</tbody>
</table>

Finally, running the simulations for a large number of iterations $K$ it can be observed that within approximately 100 iterations the estimated performance reaches almost 99% of the best found solution $M$ of the proposed algorithm.

1.3 Optimal Device-to-Device Cell Association and Load Balancing

Due to the resulting highly possible traffic congestion for some cells in a multi-cell infrastructure and the load imbalance among neighbouring areas, it is of crucial importance for the network operators to apply cooperative efficient load balancing techniques in order to provide not only satisfactory QoS for all newcomer or already active UEs but also enhance the system capacity to accommodate more users to serve.

So far, several works have been carried out to either avoid or mitigate the system's overload by injecting a variety of different offloading techniques. However, to the best of our knowledge, this is the first work that provides signalling overhead limitation and load balancing analysis for D2D paired users in cellular-based networks. The D2D links that underlay a cellular infrastructure will be mainly controlled by the network. Hence, it is mandatory to reduce the overall overhead and at the same time succeed reduced latency in controlling and orchestrating large number of D2D links in the network. Due to the booming user proliferation, the probability and need for two users to share data between each other will become more immense and lead to the irregular emergence of several D2D pairs in future networks.

In cellular networks, the most widely applied techniques for load balancing were based on the concept of borrowing channels from neighboring sparsely loaded cells such as channel borrowing without locking [16] and selective borrowing-based offloading [17] method. More recent techniques focus on the concept of traffic offloading, including cell-breathing algorithms (aiming to enhance system performance jointly with scheduling techniques [18], mobile-assisted call admission ('hot' to 'cold' cells traffic transfer [19] and overlaying ad-hoc relays integrated in cellular networks as described in [20]. It is noticed that the available schemes cannot be directly applied to traffic offloading in LTE-A Heterogeneous Networks (HetNets), since the major target here is to balance traffic load among multi-tier cells, which differ primarily in terms of physical size, maximum transmit power and so on. On that frontier, Ye et al. [21] propose a wide set of innovative user association techniques to succeed load balancing in HetNets through a number of optimization problems (which entail to be NP-hard) and their related relaxations in the form of distributed algorithms to obtain a near-optimal solution via biasing. The latter is proven to be an efficient but at the same time difficult method to be applied in real-time traffic scenarios for multi-tier heterogeneous wireless ecosystems.
Figure 5: Device-to-Device communication links in a multi-cell environment (with $h = 2d$). A number of links might cross the boundaries between neighbouring cells, hence link nodes should be connected to different BSs.

Compared with previous works, D2D communication-based load balancing techniques have the following unique features: a) D2D communications are regulated by the service-provider, including the transmission power of end users and frequency resources, b) underlaying D2D communications deploy the same frequency resources as cellular transmissions for improved spectrum efficiency [22]. Therefore, the arising levels of interference due to possible RB reuse need to be mitigated and still remains a critical challenge. This work's approach is based again on the assumption that FFR [12] is applied to restrict the inter-cell interference cases developed among the distributed users, as already specified and studied in [11]. This means that cell-edge users will be allocated specific resources from a pool depending on the respective BS to serve them. Literature concerning the in-system D2D edge-pair formation and need for load balancing among the newcomers with respect to the competition for utilizing the available spectrum is almost non-existent and is the subject of interest in this work.

In the following subsection, we provide an alternative way to examine the existence of only cell-edge D2D links that cross due to their need to communicate different serving regions in cellular networks and that will help us understand and develop an efficient optimization framework to achieve advanced BS association policies as well as signalling and load balancing improvement.

1.3.1 Scenarios, methodology and performance analysis

The probability of having both nodes of a D2D communication link been located within the geographical area of different BSs as shown in the sequel represent the well-known Buffon's Needle (BN) problem [23]. The Buffon's Needle problem examines the probability that a needle comes to lie in a position where it crosses one of the lines when dropped on a ruled two-dimensional space.
The generalization of this problem in various other surfaces is known as the Laplace extension of the baseline Buffon's problem. Assuming a hexagonal layout (coverage area of a BS) with dimension $h = 2d$ (Figure 5) and that the distance between two communicating nodes $n_1$ and $n_2$ is $l_{n_1,n_2}$, it can be shown that the probability that both nodes are within the coverage area of the same BS is as follows [23]:

$$p_0 = 1 - \frac{1}{3} \left( \frac{l_{n_1,n_2}}{h} \right)^2 - \frac{l_{n_1,n_2}}{h} \left( 4 - \sqrt{3} \frac{l_{n_1,n_2}}{h} \right) \frac{1}{\pi}. \quad (6)$$

Therefore, the probability of the nodes in the D2D link to be formed that are currently connected to a different BS will be

$$p_1 = \frac{1}{3} \left( \frac{l_{n_1,n_2}}{h} \right)^2 + \frac{l_{n_1,n_2}}{h} \left( 4 - \sqrt{3} \frac{l_{n_1,n_2}}{h} \right) \frac{1}{\pi}. \quad (7)$$

In the case where there is a multi-D2D node communication, for example a node creates D2D links with multiple nodes, then this could be considered as a polygonal with the total length being the addition of the D2D links. However, in our work we will be focusing only on the single-paired D2D connectivity type.

Figure 6: Buffon's Needle survey and Monte Carlo simulations comparison. This figure signifies the probability of having D2D link crossing within a pair of neighbouring cells. Convergence achieved for high number of iterations.

Figure 6 depicts the probability that each node in a D2D link falls within the coverage area of different BSs. Extensive simulations confirm the conceptual similarity and applicability of Buffon's Needle survey to the under study problem as, for a designated number of iterations, the output becomes...
tangent and converges to the survey's realization. The resulting probabilities were produced for varying cell dimensions as well as D2D communication range cases. The figure shows the case where each D2D link is composed of two nodes (single pair). As expected, this probability increases in proportion with the distance of the D2D communication link and decreases for cells with larger radius. In real world scenarios, the coverage areas of neighbouring cells are not well defined and therefore the percentage of D2D nodes is expected to be significantly higher in some cases. Also, in the current trend where we are witnessing further cell densification and overall decrease of the cell size radius in order to increase spatial capacity of emerging and future networks, the issue of having nodes associated with a different BS might become a significant proportion of the D2D communication links.

Load balancing techniques should be applied to distribute the increasing number of cell-edged D2D users that are not only characterized by the hitting property of the link formation but also can be formed and co-located close to the limits of all deployed cells. For the rest of the chapter we will refer to that 'hot' area of a cell as Region of Interest (RoI). Therefore, the previously depicted probability can give a good estimate of the minimum bound of D2D pairs that need to be balanced.

1.3.1.1 System model and scenario description

The under investigation D2D communication links are located close to the edge of the deployed cells and supposedly their nodes belong to a different serving geographical area, therefore are initially associated with different BSs before forming the candidate D2D pair (Figure 7a). We assume that the overall signalling overhead can be reduced by providing association of each link to a single BS, aiming to avoid any BS intercommunication and cooperation to exchange information. This approach corresponds to a more realistic networking system as it is easier to be implemented compared to multi-BS association of the users. Also, in this way, inter-BS asynchronous coordination and control inefficiency can be avoided [24]. Assuming that our topology consists of seven hexagonal cells and the total number of D2D edge users is $N_{D2D}$, the delay and overhead simplification implies pairwise association with a single BS for each constructed D2D pair and a signalling saving of cooperative flows among competing BSs, thus alleviating the need for signalling exchange equal to the number of D2D users. Thereinafter, cell-edge paired D2D users that are located in different serving areas are assumed to be regulated by a single BS, without any signalling coordination between the two BS candidates, as depicted in Figure 7b.
Figure 7: Different scenarios when nodes of a D2D link are located on cell edge and communication with two BSs is possible.

In our scenario, we assume a set of base stations $B$ and a set of D2D links $L$. Each BS $b \in B$ has a pool of remaining RBs $K_b$. Let us define by $c_{lb}$ the cost of device-to-device link $l$ connected to BS $b$, this can be considered as the average cost of connecting both nodes of a D2D link at the same BS and can be represented as follows:

$$PL_{b,n_i} = 36.7 \log_{10}(r_{b,n_i}) + 40.9 + 26 \log_{10}\left(\frac{f_c}{5}\right),$$

(8)

$$c_{lb} = \frac{PL_{b,n_1} + PL_{b,n_2}}{2},$$

(9)

where link $l$ consists of nodes $n_1$ and $n_2$, $r_{b,n_i}$ and $PL_{b,n_i}$ are the distance (in meters) and path-loss (in dB) between BS $b$ and $n_i$ for $i = 1,2$, respectively, and $f_c$ is the carrier frequency expressed in GHz. The aforementioned path-loss expression concerns urban modelling as declared in ITU-R report in [25]. Herein, fast fading impairments are not taken into consideration.

In order to view the problem of BS association and minimization of the overall cost via a mathematical programming setting, the following binary variable is defined:

$$y_{lb} = \begin{cases} 1, & \text{if link } l \text{ is connected to BS } b \\ 0, & \text{otherwise} \end{cases}$$

(10)

Based on the above, the minimum cost D2D link association to BS can be formulated as follows:
\begin{align*}
\text{minimize} & \quad \sum_{l \in L} \sum_{b \in B} c_{lb} y_{lb} \\
\text{s.t.} & \quad \sum_{b \in B_l} y_{lb} = 1, \quad \forall \ l \in L \\
& \quad \sum_{l \in L} y_{lb} \leq K_b, \quad \forall \ b \in B \\
& \quad y_{lb} \in \{0,1\}, \quad \forall \ l \in L, \forall \ b \in B,
\end{align*}

where \( B_l \) is the set of candidate competing BSs to be associated with the link \( l \). Sub-equation (12) stands for the BSs competency constraint, whereas (13) represents the capacity constraint (in terms of available resources) for each BS. The aforementioned optimization problem is an Integer Programming (IP) problem that due to the unimodular property of its inequality matrix (determinant of every square sub-matrix equals to 1) can be solved in an efficient manner, since it would have the complexity of a linear program.

Additionally, in many cases it is desirable to provide some form of load balancing between serving BSs in order to distribute the connectivity and management of the D2D links in less loaded cells so as to avoid traffic congestion and the need for RB reuse that could lead to increased interference levels among users. This problem can be formulated accordingly as follows:

\begin{align*}
\text{minimize} & \quad \sum_{b \in B} \left( \sum_{l \in L} y_{lb} \right)^2 \\
\text{s.t.} & \quad \sum_{b \in B_l} y_{lb} = 1, \quad \forall \ l \in L \\
& \quad \sum_{l \in L} y_{lb} \leq K_b, \quad \forall \ b \in B \\
& \quad y_{lb} \in \{0,1\}, \quad \forall \ l \in L, \forall \ b \in B.
\end{align*}

Note that the above is a non-linear integer optimization program which is not suitable to be solved via powerful available toolboxes on linear integer mathematical programming. However, the problem can be re-formulated as a linear problem if viewed as a max-min optimization problem which can be described as shown below:

\begin{align*}
\text{maximize} & \quad z
\end{align*}
Another way to look into the BS association would be to consider the current load level of the different candidate BSs as well as the congestion level on the fronthaul since it might be the case that D2D communication will not only incur signalling cost but also data. Therefore, if we define by $U_b$ the utilization level of BS $b$ (which might include the utilization level of the fronthaul) then we might be interested to provide load balancing by taking also into account the utilization levels of the BSs. Such an approach can be realized by solving the following mathematical program:

$$\begin{align*}
\text{minimize} & \quad \sum_{b \in B} \left( U_b - \sum_{l \in L} y_{lb} \right)^2 \\
\text{s.t.} & \quad \sum_{b \in B_l} y_{lb} = 1, \quad \forall \ l \in L \\
& \quad \sum_{l \in L} y_{lb} \leq K_b, \quad \forall \ b \in B \\
& \quad y_{lb} \in \{0,1\}, \quad \forall \ l \in L, \forall \ b \in B.
\end{align*}$$

Note again that the above formulation is a non-linear integer programming problem. In a similar manner as described before, this optimization problem can be linearized as follows:

$$\begin{align*}
\text{maximize} & \quad z \\
\text{s.t.} & \quad z \leq U_b - \sum_{l \in L} y_{lb}, \quad \forall \ b \in B \\
& \quad \sum_{b \in B_l} y_{lb} = 1, \quad \forall \ l \in L \\
& \quad \sum_{l \in L} y_{lb} \leq K_b, \quad \forall \ b \in B.
\end{align*}$$
In this paragraph, we further empower the cost minimization setting in (11), by extending the objective with the prospect of also minimizing the overall RB reuse levels. A proper RB allocation on top of a cost-efficient method for D2D connectivity could potentially entail in a reduction of the aggregate co-channel interference due to the avoidance of over-utilizing the same available resources, i.e. resource blocks. Therefore, we would like to provide cell association and RB allocation so that we also minimize the reuse of resource blocks which are currently allocated based on the FFR method. To do so, we define the following decision variable:

$$\tau_{br} = \begin{cases} 
1, & \text{if RB } r \text{ of BS } b \text{ is used} \\
0, & \text{otherwise}.
\end{cases}$$

(16)

We also define with $\rho_{br}$ an index which captures how many times a RB $r$ is used by base station $b$. Based on the above definitions, we formulate the following optimization problem which provides optimal D2D cell association and minimizes the reuse of resource blocks in the network:

$$\min \left[ \sum_{l \in L} \sum_{b \in B} c_{lb} y_{lb} ; \sum_{b \in B} \sum_{r \in R} \tau_{br} \rho_{br} \right]$$

(17)

subject to:

$$\sum_{b \in B_l} y_{lb} = 1, \quad \forall \ l \in L$$

(17a)

$$\sum_{l \in L} y_{lb} \leq K_b, \quad \forall \ b \in B$$

(17b)

$$\sum_{r \in R} \tau_{br} \leq K_b, \quad \forall \ b \in B$$

(17c)

$$\sum_{r \in R} \tau_{br} \geq \sum_{l \in L} y_{lb}, \quad \forall \ b \in B$$

(17d)

$$y_{lb}, \tau_{br} \in \{0,1\}, \quad \forall \ l \in L, \forall \ b \in B, \forall \ r \in R.$$

(17e)

In this setting, the objective functions are conflicting (i.e. cost and RB reuse varying values, respectively), hence there exists multiple number of Pareto optimal solutions. Additionally, we note that constraints (17d) are logically redundant. These constraints are implied by (17c), but are used to reduce the search effort and run-time (i.e. reducing further the search space).

1.3.1.2 Numerical investigations
In this subsection, a set of numerical investigations is detailed which shed light on the efficacy of the aforementioned optimization schemes to provide reliable BS association and load balancing performance when active communication is taking place between any constructed pair of D2D nodes within the network. A main assumption for the applicability of these problems is the existing occupation of some RBs from D2Ds that are currently being served while the newcoming and under consideration pairs will have to select from the available remaining resource pool $K_b$ of each serving BS $b \in B$ to satisfy their communication needs. Furthermore, we presume, without loss of generality, that the D2D users are uniformly distributed in the network and can be sharing the same RB with more than one users from neighboring cells as considered also in [11]. The simulation parameters are shown in Table 3.

Table 3: Simulation parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2D user distribution</td>
<td>Uniform</td>
</tr>
<tr>
<td>Region of Interest (RoI)</td>
<td>Cell-edge</td>
</tr>
<tr>
<td>Number of cells (Nc)</td>
<td>7</td>
</tr>
<tr>
<td>Cell radius (d)</td>
<td>400 m</td>
</tr>
<tr>
<td>Minimum distance for RoI consideration</td>
<td>300 m</td>
</tr>
<tr>
<td>Carrier frequency (fc)</td>
<td>2 GHz</td>
</tr>
<tr>
<td>System bandwidth (BW)</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Number of D2D investigated links (ND2D)</td>
<td>[85,160]</td>
</tr>
<tr>
<td>D2D link range</td>
<td>[20,100] m</td>
</tr>
<tr>
<td>Max number of available RBs per RoI</td>
<td>40</td>
</tr>
</tbody>
</table>

We compare the performance of the cost minimization programming problem stated in (11) with a cost-based heuristic method. The greedy heuristic allocates one RB to each D2D pair by associating each pair to the BS that provides the best channel conditions to it, without taking into consideration the total occupancy of the network but only looking into not violating the RB availability of the respective cell. The algorithm runs in a centralized manner by sequentially assigning RBs to the involved users and when the resource buffer of the related BS to serve is full, the transmission needs of the respective users is satisfied from the competing BS candidate. The cost minimization difference is expected to be somehow low for the two compared techniques since the objective for mitigating the overall cost and achieving an efficient RB assignment combination is implemented in a similar way.

Figure 8 highlights this slight differentiation and reveals the increasing mean gain/reduction proportionally with the increase of utilization levels and traffic congestion in relation to the remaining RBs to be occupied. This proves that the cost minimization IP formulation performs better in congested scenarios. In other words, the gains are achieved when really needed, i.e. in congested epochs which are envisioned to be of frequent occurrence due to the exponential growth of user emergence in current and future networks. The results are normalized for ease of comprehension and are depicted for different levels of congestion. For these simulations highly utilized BSs were assumed as for low congestion scenarios the two compared methods have similar outputs. The highest cost minimization gain is met on the most bottlenecked case (congestion level > 90%) in a percentage of almost 3.15%.
Figure 8: Normalized cost comparison for cost minimization IP solver and Cost-based Heuristic (CbH) algorithm for high congestion levels.

The aforementioned notable but also small gain is the reason we enhance the cost minimization problem by introducing an additive sub-objective to efficiently control the unassigned users that need to be orchestrated. As implied in problem (17), this programming solver looks into allocating lightly used RBs of the candidate BS to the unallocated D2D links given that all involved BSs are active and highly utilized. Furthermore, due to the limited RB availability and the huge number of users, it can be easily conceivable that several resources will be over-utilized. To this direction, we apply this solver to high congestion instances where the resources are reused within the network. Figure 9 proves that for different traffic cases and occupancy of the network, this setting achieves better performance compared with an average utilization agnostic method that assigns resources randomly picked from the available pool of RBs. On the other hand, the worst case would be to assign users to satisfy their transmission needs by using the over-utilized resources. As depicted, the optimal case of utilizing the less used RBs outperforms the worst case and average utilization agnostic case in a percentage of 45% and almost 22% over-utilization avoidance gain, respectively. Similar behavior is observed for more congested instances (> 70%) where the achieved gain remains significant.
In regard to the problem of load balancing within the deployed topology from the perspective of D2D users, we compare the performance of the proposed load balancing IP frame (problems (15), (19)) with two heuristic algorithms. The first one, namely the cost-based heuristic (CbH), was described afore in the cost minimization problem. The second greedy heuristic is a load-based heuristic (LbH). LbH bases its allocation rationale on looking into the current load of the network and each cell-edge's capacity in order to assign resources to the less loaded between competing cells. The latter theoretically works better in the case of relatively balanced cells but because of its dependency to load fluctuation scenarios (varying RB availability) can be slightly weakened compared to the former in initially load imbalanced cases. Also, it is obvious that the LbH method outperforms the CbH when the objective is to balance the load of the deployed cells.

Figure 9: Normalized utilization balancing comparison for optimal allocation of RBs.

Figure 10 shows a statistical significance and supremacy of the proposed optimization problems compared to both CbH and LbH algorithms for a sample of ten different deployment scenarios. The results derived from a round of independent simulations and for this sampled cases the standard deviation $\sigma$ of the CbH and the LbH methods is almost 3.5 and 2.7 times bigger, respectively, compared to the LB-IP solution. This signifies less spread and more balanced assignment of the available RBs to the newcoming D2D users by using the optimization framework. The least observed deviation of the LB-IP solution to those algorithms (scenario 4 in Figure 10) is characterized by a highly restricted scenario where the distribution of users is very similar for all cells and the number of available RBs is limited and close to the number of edge D2D pairs.

In this case, 7% and 13.6% standard deviation percentage differentiation characterizes the result of CbH and LbH respectively compared to the optimal solution.
Figure 10: Simulation instances of Load Balancing (LB) optimization solver, Load-based (LbH) and Cost-based heuristic (CbH) algorithmic solutions.

A study was conducted for varying number of D2D cell-edge links (bounded accordingly as defined in Figure 6 for a total number of formed D2D pairs) where the mean standard deviation is the metric for the results shown in Figure 11. This figure reveals the consistency of the proposed optimization problem as a solver for both light and more congested ecosystems where the availability of resources is more likely to become constrained by the uprising number of users. The IP solution outperforms in terms of mean standard deviation the LbH as well as the CbH load balancing algorithms in a percentage of almost 44% and more than 53% difference respectively, as depicted below. The increasing tendency of the LB-IP solution can be justified by the fact that, while retaining the constrained $K_b$ capacity values, some of the newcoming D2Ds might fully utilize the available resources on a specific cell; therefore, the remaining unassigned D2D links should be coupled with the related BS that has available resources. This entails in increasing association variance for increased number of users. Despite the random deployment of D2D links for different cases, the results highlight the performance gain of the LB-IP solver per case. In addition, after running the LB-IP solution a large number of iterations by retaining the same number of randomly distributed D2Ds (115 users for each instance) but for different deployment scenarios, the average elapsed time was estimated on 0.2517 seconds and could converge at least one order of magnitude faster by using hardware accelerating techniques to improve its time-running performance.
Figure 11: Mean standard deviation estimations for Load Balancing (LB-IP) solver, Load-based (LbH) and Cost-based heuristic (CbH).

Finally, a set of results is presented for problems (14) (basically, its linear relaxation in (15)) by including and examining the effect of the utilization level $U_b$ of each cell. The results depicted in Figure 12 were produced for randomly generated utilization levels for each BS without violating any resource constraint. The evaluation of the included load balancing schemes is likewise compared to the figure above and the differences are mainly arithmetic and incidental due to the randomness of the parameter values for different scenarios and the involvement of the $U_b$ limitations. The optimal recorded mean standard deviation values differ in a percentage of close to 50.5% and 55.3% compared to LbH and CbH algorithmic outputs, respectively.
1.4 Future research planning

Considering the latest results on load balancing for cell-edged D2D links, a very important issue that needs to be taken into consideration is the interference employed due to the possible existence of multiple D2D pairs in each cell’s ambiguity (RoI) coverage area. According to the applied FFR method, a cell-edge or crossing D2D pair that associates with a BS can mainly cause interference to cellular UEs in adjacent outer cell regions that might utilize the same resource. The D2D transmitter generates an interference range, where the increased number of users located within its transmission range enhances the probability of a strong co-channel interference. Especially in dense networks, the need to provisionally avoid immense interference scenarios is obvious. By applying a joint optimization framework that includes not only a load balancing criterion but also interference-aware intelligence, the users’ reduced interference levels could entail better achievable system throughput and could lead to an overall performance improvement. Further research will be applied to this direction and expected results will include throughput analysis and resource utilization improving conditions.
2. Algorithms for SONs and cognition in LTE-A for handovers optimization

2.1 Introduction

An increase in the number of connected mobile devices as well as the introduction of smart devices with higher computational power, larger capacity and ability to support demanding services is expected to result in a tenfold increase in the mobile data traffic over the next four years [1]. This trend is clear in (Figure 14) where what should be underlined is the small percentage of traditional voice services and the lion’s share of market that data demands (video demanding, VoIP services, internet navigation) will possess. The main problem that cellular networks will face is how to serve this amount of traffic ensuring the quality of service (QoS) and quality of experience (QoE) for the users.

Note that, the spectrum allocated to cellular operators is limited and, consequently, an efficient utilization of the existing resources emerges as a need to meet the ever increasing traffic demand. Last studies suggest that the densification of the radio access network (RAN) is one of the key aspects that could provide higher capacities, and particularly the extensive deployment of small cells (i.e. femto cells, metro cells, pico cells and micro cells) [27][28].

These nodes can operate on the same or different frequencies from the traditional macro cells, their operational cost is low in both energy and financial dimension and their installation does not require extreme effort. Their coverage area is smaller than the macro cells and last but not least can offer higher spectrum efficiency. The relationship between the macro and the small cells spectrum efficiency can be found in [29]. On average the spectrum efficiency of small cells could be 25% better from the corresponding macro site due to the reduced interference among the small cell tier.
The presence of cells with varied features in the same area will have as a consequence the transformation of the traditional macro only deployments to the so-called heterogeneous networks (HetNets). This densification of cellular networks will emerge the need for further standardization of procedures and proposal of new algorithms, since techniques used in macro only scenarios cannot meet the demands of the new architecture in an efficient way. For example mobility management (MM) is an attractive concept for both academia and industries given that the efficient management of moving users (user equipment UE) is extremely challenging and requires a lot of effort from operators. In this environment self-organized networks (SON) arise as the necessary capability for the network to deal with the increasing demands for different type of services, the complexity of future wireless access networks and the need to reduce costs from a market perspective. In alignment with the aforementioned mobility and SON challenges, short-range communications will be a key issue in future cellular networks.

This section is focused on the analysis and design of efficient Radio Resource Management (RRM) algorithms in the framework of SONs. In particular MM solutions aiming at improving the performance of the handover (HO) process in challenging heterogeneous networks will be developed. Furthermore, the future work which will be dedicated to the study of short-range communications (Device-to-Device D2D) and to proposals that enhance system's performance and limits will be presented. The main emphasis lies on providing solutions and techniques that will permit a scale free and flexible deployment of future cellular networks in respect to the QoS and QoE of users.

### 2.2 Handover procedures

Mobility Management (MM) is one of the cornerstones in HetNets, since it is severely affected by the aforementioned rapid changes in networks' architecture. MM is a wide concept containing procedures like cell identification, access control, cell search etc. [30] that will be confronted with adversities due to the extensive introduction of small cells. Handover (HO) procedure, which is vital for the proper operation of the cellular networks to guarantee continuous and seamless connection of the users [31], will also have to be re-considered to overcome the inefficiency of existing algorithms and techniques to deal with the dense nature of future networks. Thus, the necessity for more sophisticated HO solutions to address problems arisen from the integration of conventional HO algorithms into HetNets, still remains as an open issue [32].

![Figure 14: Exabytes per Month of Mobile Data Traffic by 2019](image)
In this framework, study of handover (HO) has attracted large attention from both academia and industry [33]. It has been proven that using in HetNets the same set of HO parameters (i.e Hysteresis Margin (H), Time-to-Trigger (TTT)) as in homogeneous networks degrades HO performance. Specifically, reduced coverage areas of small cells may result in frequent HOs, Ping-Pongs (PPs), Handover Failures (HOFs) and Radio Link Failures (RLFs), mainly for fast User Equipments (UEs) [32]. Therefore, the aforementioned transformation undergone in the networks’ architecture imposes a reconsideration of the HO procedure to keep up with the increasing complexity. Given that, as stated above, the degradation of the HO metrics is tightly coupled with the UE profile (e.g. the UE speed), the proposed solutions must be UE specific and in alignment with the Self-Organized Networks (SON) concept.

3GPP suggests in [32] three important metrics to evaluate the HO performance and, consequently, the impact of the new architectures: Handover Failure (HOF), Ping-Pong (PP) and Radio Link Failure (RLF) probabilities.

HOF is a term used to describe the loss of radio link connection during the HO process and can occur in three cases. First, when the RLF timer (i.e. T310) is still running at the end of the HO preparation time. This leads to the existence of bad radio connection between the UE and the source cell that might result in unsuccessful delivery of whether the UE measurement report or the HO command. The second case occurs when T310 expires while TTT is still running. Finally, in the last case the received signal quality from the target cell at the end of HO execution time is below a threshold, namely $Q_{out}$ [32].

With regard to PP, UEs moving through a small cell, and particularly fast users, are prone to short sojourn times in the small cell, known as Time of Stay (ToS). When this time is below the Minimum Time of Stay (MTS), the HO pair is defined as a PP. The value recommended by 3GPP for MTS is 1s [32].

Finally, RLF is caused by the degradation of the signal quality from the serving cell. Specifically, the RLF occurs when the Signal to Interference Noise Ratio (SINR) falls below $Q_{out}$ and, subsequently, it remains below $Q_{in}$ during 1s. The typical values used in the literature for $Q_{out}$ and $Q_{in}$ are -8dB and -6dB, respectively [32].

In this context, where networks will be characterized by their heterogeneity, the appropriate selection of the HO parameters will be instrumental to guarantee their efficient operation and acceptable rates of HOF, PP and RLF.

Solutions to the problem of frequent HOs in HetNets were investigated in a series of proposals aiming at adapting the Hysteresis Margin according to network conditions and UE state [34]-[37]. For instance, the work in [36] proposes an adaptive hysteresis algorithm that accounts for factors like UE speed, type of service and the load difference between the target and the serving cells, in order to reduce HOFs in a two-tier network. Similarly, an efficient handoff algorithm is presented in [37]. The idea is to combine both signals from the macro and the small cell, to compensate the uneven transmission power between nodes in different tiers, thereby creating an adaptive hysteresis margin and encouraging HOs to small cells. However these proposals did not take into consideration a key HO parameter: the Time to Trigger (TTT).

The importance of TTT on the mobility performance in HetNets is studied in [38], where a solution to mitigate the undesired high rates of RLFs and PPs is presented. The proposed scheme is based on the optimization of the TTT parameter according to the speed of the UEs in a scenario with macro and pico cells. In [38] and [39] the objective is to limit the RLF per HO attempt to 2%, since such limit should guarantee acceptable QoS.
A study on mobility performance in co-channel HetNets is also presented in [40] where, in addition to [38], the relationship between small cells' size and the proper TTT value is stressed out. In more detail, the problem of performance degradation, in terms of increased HOF, RLF and PP probabilities, when a HO takes place between different type of cells (Macro, Pico) is analyzed. It is concluded that using the same TTT for both macro and pico cells is not effective. Therefore, different TTT values are proposed for different types of cells with respect to UEs' speed (slow, fast). An attempt for further improvement of the HO procedure is presented in [41]. The proposed policy is based on a cell-pair TTT selection taking also into account the speed of the UEs. The rationale behind this proposal is to prevent fast UEs from performing frequent HOs to small cells, unless it is necessary due to severe interference, and allow the offloading of low speed users' traffic to small cells tier. Results reveal that TTT must be increased for high mobility users in macro to pico HOs.

The strong relationship between HOF and PP is emphasized in most of the related works. This relationship could be summarized as follows: the delay of the HO (by increasing the TTT), despite preventing UEs from unnecessary HOs (i.e. PPs), might result in higher HOF rate; on the contrary, if HO is encouraged through low TTT values, HOF rate is reduced due to fast HO execution, but it leads to an increase of the number of PPs. The dependence of the handover performance on different ratios, e.g. HOF, PP and RLF, led Jansen et al. to define a handover performance metric for evaluation purposes in [42]. The new metric is a linear combination of HOF, PP and Call dropping ratio, and the weight of each ratio is calculated iteratively.

Although the relationship between HO metrics is outlined in most of the aforementioned proposals, the absence of solutions based on mathematical analysis is evident. Motivated by the above, the work in [43] uses a geometrical approach to derive closed-form expressions for HOF and PP probability as a function of UE speed, TTT and range expansion bias. The conclusion is that there is an optimal TTT value for a given UE and cell.

The geometric approach followed in [43] as well as the assumption regarding the circular area of the small cell [32], were used in [44] for expressing the HOF probability in a HetNet deployment as a function of the L3 sampling period. Finally, a work that investigates the dependency of the outbound HOF probability on parameters like UE speed, TTT and shadowed channel fading can be found in [45]. Closed-form expressions of the HOF probability as a function of the aforementioned parameters are derived by using stochastic geometry.

Although the existing literature has pointed out the importance of the main parameters in HO performance (e.g. the type of cells involved in the HO and the speed of the UEs), there is a parameter that has been completely overlooked: the distance between the target and the source cells (i.e inter-site distance) in a macro to small cell HO. Distance between the target and the source cells determines signal and interference levels. Based on this, the range of the cells, particularly the small cells', is also connected to this inter-site distance. As a consequence, metrics such as HOF, RLF and PP probabilities will be affected by the location of the small cells within the macro cell coverage area.

### 2.3 Handover Performance in LTE-A HetNets through Inter-Site Distance Differentiation
#### 2.3.1 Scenario and Methodology

In this first work we study the impact of the distance between the center of the macro and small cell on HO performance, and how TTT should be selected to improve the system performance. Our aim is to show that each deployment should be treated differently in order to reduce frequent HOs while ensuring uninterrupted service of UEs.
The scenario under study consists of a small cell located a distance $D$ from the center of the overlaid macro cell in a two-tier deployment, as depicted in Figure 15. In this scenario UEs cross the small cell and execute a HO (or not) based on the HO procedure definition. Regarding the HO decision, it is based on the A3 event [31]. According to this, a UE triggers the HO procedure from a source cell to a target cell if the Received Signal Strength (RSS) from the latter, also known as Reference Signal Received Power (RSRP), is higher than the RSS from the former plus the Hysteresis margin. Note that we consider the path loss model included in Table 4, while the impact of fading and shadowing is neglected. This condition should hold for a $TTT$ period before the HO is executed. Thus,

$$RSS_t \geq RSS_s + H$$  \hspace{1cm} (18)

where $RSS_t$ and $RSS_s$ are the Received Signal Strength received from the target and the source cells, respectively, and $H$ is the hysteresis margin (all of them expressed in dB). After appropriate manipulation, (18) may be rewritten as:

$$P_{Tt} - A_{ot} - a_t 10 \log_{10} d_t \geq P_{Ts} - A_{os} - a_s 10 \log_{10} d_s + H$$  \hspace{1cm} (19)

Figure 15: Macro-Small cell HO scenario

where $P_{Tt}$ and $P_{Ts}$ are the transmitted power by the target and the source cells, $A_{ot}$ and $A_{os}$ are the distance independent components of the path loss models and $a_t$ and $a_s$ are the exponents of the corresponding path loss models (all of them expressed in dB). Reformulating (19) results in

$$d_t \leq d_s 10^\frac{H}{10}$$  \hspace{1cm} (20)

where $d_t$ and $d_s$ are the distances to the target and source cells, respectively.
\[
\gamma = \frac{(P_{Tt} - P_{T_s}) - (A_{ot} - A_{os})}{10} \tag{21}
\]

For a generic UE \(i\) located a distance \(d_t\) from the small cell (i.e. the target cell) and \(d_s\) from the macrocell (i.e. the source cell) with an angle \(\Theta_i\) (see Figure 15) the relationship between \(d_t\) and \(d_s\) is described by,

\[
d_s = \sqrt{D^2 + 2Dd_t\cos(\Theta_i) + d_t^2} \tag{22}
\]

and consequently

\[
d_t \leq 10^{-\frac{\gamma-0.1H}{\alpha_s}} \left( D^2 + 2Dd_t\cos(\Theta_i) + d_s^2 \right)^{\frac{\alpha_s}{2\alpha_t}} \tag{23}
\]

After convenient calculations, and assuming \(\alpha_s = \alpha_t\), (23) may be written as

\[
d_t \leq R_1 \tag{24}
\]

with

\[
R_1 \approx \left( \frac{D}{10^{-\frac{\gamma-0.1H}{\alpha_s}} - 1} \right) \left( \cos(\Theta_i) \pm \sqrt{\cos(\Theta_i)^2 + 10^{-\frac{2(\gamma-0.1H)}{\alpha_s}} - 1} \right) \tag{25}
\]

As \(R_1\) is positive by definition, \(\varphi \geq 0\), and so it can be concluded that \(R_1\) rises as \(D\) is increased.

The region where (24) holds is defined as the coverage area of the small cell and it is coloured in blue in Figure 15. In the boundary of this region (i.e. \(d_t = R_1\)) the \(TTT\) counter of the inbound handovers starts, and the HO must be necessarily executed within this region, i.e. \(d_t \leq R_1\). Otherwise, the \(TTT\) counter is stopped and the HO canceled. Yet, as pointed out previously, a HOF occurs if the \(TTT\) counter expires within this region but the \(SINR\) from the source cell is below \(Q_{out}\) (i.e. \(SINR < Q_{out}\)) [46]. This region is coloured in red in Figure 15 and is defined, similarly to (24) and (25), by the following expression:

\[
d_t < R_2 \tag{26}
\]

where

\[
R_2 \approx \left( \frac{D}{10^{-\frac{-2(\gamma+4Q_{out}-N)}{\alpha_s}} - 1} \right) \left( \cos(\Theta_i) \pm \sqrt{\cos(\Theta_i)^2 + 10^{-\frac{-2(\gamma+4Q_{out}-N)}{\alpha_s}} - 1} \right) \tag{27}
\]
where $N$ is the downlink noise power expressed in dB. It is worth noting that both regions, defined by $R_1$ and $R_2$, depend on the inter-site distance ($D$) and the entry angle of the user $\Theta$. Therefore, the ToS of the UEs, tightly coupled with $R_1$, $R_2$, the speed of the UE, and the TTT value, depends on $D$ as well. In the sequel, an exhaustive analysis of the impact of $D$ on the HO performance is presented.

### 2.3.2 Numerical Results

The dependence of the HO performance is evaluated through extensive simulations in the urban scenario defined in [47],[48]. The simulated scenario consists of a small cell and a macro cell with an inter-site distance $D$, ranging from 40m to 240m. The first interference ring composed of 6 macro cells has been also simulated to include the co-channel interference of the macro cells tier. UEs move at 60 km/h, since they are regarded as high speed users for urban scenarios.

**Table 4 Simulation Parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISD for Macro cell</td>
<td>500 m</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Macro and Small Cell Frequency</td>
<td>2 GHz</td>
</tr>
<tr>
<td>Macro cell Path-Loss</td>
<td>$128.1 + 37.6 \log_{10}(\text{distance})$, {distance in km}</td>
</tr>
<tr>
<td>Small cell Path-Loss</td>
<td>$140.7 + 36.7 \log_{10}(\text{distance})$, {distance in km}</td>
</tr>
<tr>
<td>Macro cell transmitted power</td>
<td>46 dBm</td>
</tr>
<tr>
<td>Small cell transmitted power</td>
<td>20 dBm</td>
</tr>
<tr>
<td>HO A3 Hysteresis Margin</td>
<td>3 dB</td>
</tr>
<tr>
<td>TTT values</td>
<td>128, 160, 320, 640 ms</td>
</tr>
<tr>
<td>HO Preparation Time</td>
<td>50 ms</td>
</tr>
<tr>
<td>HO Execution Time</td>
<td>40 ms</td>
</tr>
</tbody>
</table>

The wide range of $TTT$ values defined in [31],[38] has been considered and simulated, but only illustrative values have been included in the subsequent figures for the sake of clarity. The rest of the parameters are listed in Table 4.

The metrics used to evaluate the simulations’ performance are the aforementioned $RLF$ ($P_{RLF}$), $PP$ ($P_{PP}$) and $HOF$ ($P_{HOF}$) probabilities. However, it may also be evaluated in terms of the probability of performing a HO ($P_{HO}$) and the probability of not performing a HO ($P_{NHO}$). In that sense, a UE entering the coverage area of a small cell may perform a HO (no matter if it is a PP or not), suffer from RLF, or not performing a HO (due to a HOF or to a high TTT value). The relationship between these probabilities is given by:

$$P_{NHO} = 1 - P_{HO} - P_{RLF}$$  \hspace{1cm} (28)

With the objective to analyze the different trends of the results as a function of $D$, and based on the figures to be presented in this paragraph, four sets of scenarios are defined: $A_1$, $A_2$, $A_3$ and $A_4$. 
Specifically, in the first set of scenarios ($A_1$) the small cell is deployed close to the eNodeB, i.e. 40 m $\leq D < 100$ m. Notice that $D < 40$ m is not considered since, as a consequence of the unbalanced transmitted power of the small cell and the macro cell (20 dBm and 46 dBm respectively), the size of the coverage area of the small cell would become too small. Likewise, $A_2$ is characterized by 100m $\leq D < 160$m, $A_3$ by 160m $\leq D < 170$m, and finally $A_4$ by 170m $\leq D < 240$m.

In general terms, the HO is performed when (18) holds for a $TTT$ period. Thus, for a user moving at speed $v$, in principle it will perform a HO if its trajectory within the area defined by $R_1$ is lengthier than $v \cdot TTT$. Taking into account that the coverage area of the small cell (i.e. $R_1$) grows with $D$, as shown in (21), the probability of performing a HO ($P_{HO}$) should grow when $D$ is increased and should fall when $v$ grows. For the same reason, the Ping-Pong probability ($P_{pp}$) should be low when $R_1$ is small (since $v \cdot TTT$ is higher or comparable with the coverage area diameter), higher for intermediate $R_1$ values, and so $R_1$ reaches its highest values.

Likewise, the HOF region defined by $R_2$ also grows with $D$ and, consequently, the HOF probability ($P_{HOF}$) should rise as well. All these counteracting factors make the appropriate selection of $TTT$ one of the key aspects to improve the HO performance. In the sequel, the effect of these opposite trends, which is translated into the aforementioned metrics, is analyzed in Figure 16 - Figure 20.

Figure 16 and Figure 17 show the $P_{NHO}$ and the $P_{HO}$ of a UE moving at 60 km/h. As pointed out in (28), $P_{HO}$ and $P_{NHO}$ are complementary (i.e. $P_{NHO}=1-P_{HO}$) when $P_{RLF}=0$. Although the probability of RLF will be analyzed later, it is worth noting that it is equal to 0 in $A_1$ and $A_2$, and so the analysis of $P_{HO}$ and $P_{NHO}$ is equivalent for $D<160$m. It may be observed in Figure 17 that, as $R_1$ grows with $D$, the $P_{HO}$ also rises when the small cell is deployed far from the macro cell center. Yet, it may be seen that in $A_1$ and $A_2$ there is a maximum $P_{HO}$ for $TTT$ values. This maximum is caused by two factors: firstly, the increase of $R_1$ makes the $TTT$ counter more likely to expire within the coverage area of the small cell, and so $P_{HO}$ grows; secondly, as $R_2$ grows as well, the probability of having a HOF also rises. According to Figure 17, when $D>50$m for $TTT=320$ms (and $D > 90$m for $TTT=640$ms) the number of HOFs compensates the increase of $R_1$ and so it results in a fall of $P_{HO}$. This phenomenon does not happen for low $TTT$ values, since in this case the $P_{HOF}$ is high even for small $D$ values.

Regarding the scenarios with a distant small cell (i.e. $A_3$ and $A_4$), the $P_{HO}$ maintains the upward trend as a consequence of the increase of the small cell coverage area. However, high $D$ and $R_1$ values are translated into a decrease of the $SINR$ and the rise of $P_{RLF}$. This increase of $P_{RLF}$ limits the $P_{HO}$ in $A_4$.

Based on the presented results, the HO performance is highly dependent on the size of the small cell coverage area ($R_2$) and the size of the HOF region ($R_2$). Yet, the impact of $R_1$ and $R_2$ is also connected to the distance covered by the user before the $TTT$ counter expires. In other words, the distance $v \cdot TTT$ determines whether $R_1$ and $R_2$ values are big or not. This is the reason why, for a given $v$, the $P_{HO}$ is higher for small $TTT$ values.

As mentioned, the RLF occurs when the $SINR$ received from the source cell remains below $Q_{in}$ longer than 1s. This situation can only occur when the coverage area of the small cell ($R_1$) is big enough. In Figure 18 the size of the small cell coverage area is only big enough when $D>160$m. Note that the $SINR$ of the source cell has two main components: on the one hand the $RSS$ received from the source cell (i.e. the macro cell) and on the other hand the interference from the target cell (i.e. the small cell). Both components depend on the distance to the source cell ($D - R_1$) and the distance to the target cell ($R_1$). For low $TTT$ values, RLF events are declared only for a small range of entry angles. Therefore, and as shown in Figure 18, in these cases the value of the $SINR$ at a distance $D - R_1$ from the macro cell and $R_1$ from the small cell is particularly important. Specifically, if $D$ grows faster than $R_1$, the RLF will grow. On the contrary, if $R_1$ grows more than $D$, the RLF will decrease. This trade-off may be observed in $A_4$ of Figure 18 with $TTT=128$ms and $TTT=160$ms. Specifically, while $D$ gradually increases, $R_1$ grows...
more than $D$. However, closer to the macro cell edge, where the interference from the adjacent macro cell is stronger, $D$ is the one that grows faster than $R_1$. This can be depicted in Figure 18 with the upward trend for $D \geq 230m$.

The behavior of the PP probability (Figure 19) is also related to $R_1$ and $v \cdot TTT$. Thus, when $R_1$ is much smaller than $v \cdot TTT$, the $P_{PP}$ tends to 0 (the UEs do not handoff); when $R_1$ is comparable with $v \cdot TTT$, the $P_{PP}$ rises; finally, when $R_1$ is much larger than $v \cdot TTT$, the $P_{PP}$ drops (since UEs handoff but their ToS is long).

Finally, the $P_{HOF}$ is plotted in Figure 20. As explained previously, when the $TTT$ counter expires within the HOF region defined by $R_2$, a HOF occurs. Since the $TTT$ counter is triggered at a distance $R_1$ from the small cell and the HOF starts at a distance $R_2$ from the small cell, the $P_{HOF}$ depends on the relationship of the distance $R_1-R_2$ and the distance $v \cdot TTT$. For instance, if a UE is moving from the macro cell center to the small cell center, the UE will suffer from a HOF if $R_1 - R_2 \leq v \cdot TTT \leq R_1 + R_2$. As shown in Figure 20, this results in high $P_{HOF}$ for small cells close to the macro cell center when $TTT$ values are low, and a decrease of the $P_{HOF}$ as the small cell is moved away from the macro cell. On the contrary, and for the same reason, the trend is the opposite for high $TTT$ values.

Based on the presented results, it is clear that in HetNets the $TTT$ value should be selected in a cell-pair basis. Specifically, low $TTT$ values are appropriate for small cells deployed far from the macro cell center. By doing so, $P_{PP}$, $P_{RLF}$ and $P_{HOF}$ are reduced while the usage of the small cells is maximized (i.e. high $P_{HO}$). Conversely, for small cells deployed in the vicinity of the macro cell center, high $TTT$ values are more convenient.

2.3.3 Performance results

The problem of mobility management in HetNets, and particularly the selection of the $TTT$ value, has been addressed. It has been proved through extensive simulations that the performance of the systems in terms of HO, RLF, PP and HOF probabilities, is tightly coupled with the distance between the small cell and the macro cell, and the speed at which the UE is moving. Hence, it has been shown that high $TTT$ values are appropriate for small inter-site distances, while lower $TTT$ values provide better results for small cells deployed in the edge of the macro cell [49].

![Figure 16: No HO probability, speed = 60 km/h](image-url)
Figure 17: inbound HO probability, speed = 60 km/h
Figure 18: RLF probability, speed = 60 km/h

Figure 19: PP probability, speed = 60 km/h
2.4 The Impact of Inter-Site Distance and Time-to-Trigger on Handover Performance in LTE-A HetNets

The aim of our second contribution on Handover procedures, is firstly to prove the dependency of the HO performance on the inter-site distance in a two-tier network. Secondly, to derive closed-form expressions for the different HO performance metrics as a function of inter-site distance and speed of UEs. Finally we aim to demonstrate how an appropriate TTT selection should be made, taking into account the inter-site distance and the UE's speed in order to offload traffic to small cells without degrading HO performance and ensuring that the RLF probability will not exceed 2% per HO as suggested in [38] and [39].

2.4.1 Scenario and Methodology

HO performance is studied in a two tier deployment consisting of a small cell located at a distance $D$ (inter-site distance) from the center of the overlaid macro cell. UEs connected to the macro cell, cross the area of the small cell, moving on a straight line triggering a HO. This procedure is based on the A3 event, according to which a HO initiates when the received signal strength ($RSS$) from the target cell becomes an offset better than the source cell for a period equal to $TTT$ [31]. The offset is known as Hysteresis Margin. The above definition can be expressed as follows

$$RSS_t \geq RSS_s + H$$ (29)

where $RSS_t$ and $RSS_s$ are the Received Signal Strength received from the target and the source cells, respectively, expressed in dBm, while $H$ stands for the Hysteresis Margin, expressed in dB. Note that
the region where (29) holds, defines the coverage area of the small cell. As suggested in [32],[43] and [45], the region can be approximated as a circumference centered at the small cell site, depicted in Figure 21 as the light shaded circle. Given this scenario, the HO performance is expressed in terms of HOFs, RLFs and PPs probabilities [32].

Let us consider a generic UE located at the boundary of the small cell, defined in (29), at distance \(d_t\) from the target cell, and \(d_s\) from the source cell (both of them expressed in km).

The relationship between \(d_s\) and \(d_t\) is then described by

\[
d_s = \sqrt{D^2 + 2Dd_t \cos(\Theta_{inner}) + d_t^2}
\]  

(30)

where \(\Theta_{inner}\) is defined as the angle formed by the straight line joining the UE location and the small cell site, and the line joining the small cell and the macro cell sites (see Figure 21). Focusing on (29), it can be reformulated as

\[
d_t^{a_t} \leq d_s^{a_s} 10^{\frac{\gamma - H}{10}}
\]  

(31)

where \(a_t\) and \(a_s\) are the exponents of the corresponding path loss models and

\[\gamma = \frac{(P_{T_t} - P_{T_s}) - (A_{s_1} - P_{t_1})}{10}\]

where \(A_{s_1}\) and \(A_{t_1}\) are distance independent components of the path loss models [47],[48], while \(P_{T_t}\) and \(P_{T_s}\) stand for the transmitted power from target and source cell respectively.

If we define the small cell radius (\(R_\Theta\)) as the maximum \(d_t\) (for a given \(\Theta_{inner}\)) for which (31) holds, the radius may be numerically calculated from the following expression

\[R_\Theta = 10^{\frac{\gamma - 0.1H}{a_t}} (D^2 + 2R_\Theta D \cos(\Theta_{inner}) + R_\Theta)^{\frac{a_s}{2a_t}}
\]

(32)

At the edge of the region defined by (29), (31) or (32), \(TTT\) initiates and should expire within it in order for a HO to be executed successfully. However, if at expiration of \(TTT\), the UE is located inside the small cell but the SINR received from the source cell is below \(Q_{out}\), there is a HOF. This region, where \(SINR_s \leq Q_{out}\), is depicted in Figure 21. as a white coloured circle (the inner circle, also known as HOF circle) and has a radius hereof denoted as \(r_\Theta\). Similarly to (32), \(r_\Theta\) is given by

\[r_\Theta = 10^{\frac{\gamma + 0.1Q_{out}}{a_t}} (D^2 + 2r_\Theta D \cos(\Theta_{inner}) + r_\Theta)^{\frac{a_s}{2a_t}}
\]

(33)

Once a UE is associated to the small cell, the outbound HO procedure starts (i.e. the \(TTT\) countdown) when (1) holds. Note, however, that in this case the macro cell plays the role of target cell and the small cell is the source cell. Following the same reasoning, the radius of the outbound HO, namely \(S_\Theta\) (the dark shaded circle in Figure 21) may be calculated numerically from the following

\[S_\Theta = 10^{\frac{\gamma + 0.1H_t}{a_s}} (D^2 + 2S_\Theta D \cos(\Theta_{inner}) + S_\Theta)^{\frac{a_s}{2a_t}}
\]

(34)
where $H'$ is the hysteresis margin used to handoff from the small cell to the macro cell. It is worth noting that all these radii, i.e. $R_\Theta$, $r_\Theta$ and $S_\Theta$, do depend on the inter-site distance $D$ and $\Theta_{inner}$. As UEs are assumed to be distributed randomly around the scenario, the mean radius $R$ can be expressed as

$$R = \mathbb{E}[R_\Theta] = \frac{1}{\pi} \int_0^\pi R_\Theta d\Theta_{inner}$$

(35)

Likewise, the mean of $r_\Theta$ and $S_\Theta$ are given by $r = \mathbb{E}[r_\Theta]$ and $r = \mathbb{E}[S_\Theta]$, respectively.

In 2.4.2, the analytical expressions for inbound HO ($P_{HO}$), RLF ($P_{RLF}$), HOF ($P_{HOF}$) and PP ($P_{PP}$) probabilities are derived as a function of $R$, $r$ and $S$ to further analyze the impact of UE's speed and $D$ on the appropriate selection of TTT values. Although there are other factors that could impact the performance of the described probabilities, e.g. the L3 filter or the measurement errors [44], the presented model will assume perfect measurement and no L3 delay to focus on the effect of the TTT and inter-site distance.

### Figure 21: Small cell area and considered angles

**2.4.2 Mathematical Analysis**

Let us focus on the same generic UE, located at distance $R$ from the center of the small cell. This UE, moving at speed $v$, crosses the small cell coverage area with an entry angle $\theta_e$ (see Figure 21). Due to symmetry, the whole analysis will be done hereafter for $0 \leq \theta_e \leq \frac{\pi}{2}$. In this situation, this UE may
perform a HO, suffer from a HOF, an RLF, or/and a PP. Before proceeding with the analysis of the aforementioned probabilities, it is necessary to define a set of important angles.

First, let \((TTT + T_p)\) be the time needed for an inbound HO to be completed, from now on denoted as \(T\). The distance covered by the UE during time \(T\) is then equal to \(vT\). Accordingly, if we define \(\theta_i\) as the maximum entry angle for which the inbound HO can be completed (i.e. the distance covered within the small cell coverage area is, at least, \(vT\), as shown in Figure 21), it may be expressed as

\[
\theta_i = \arccos \left( \frac{vT}{2R} \right)
\]

(36)

with \(0 \leq v \leq \frac{2R}{T}\). If \(v > \frac{2R}{T}\), the UE will not handoff to the small cell regardless of the entry angle.

The second angle to consider, \(\theta_t\), is defined as the maximum angle for which \(T\) expires within the HOF circle (Figure 21). By using the Cosine Law, \(\theta_t\) may be expressed as

\[
\theta_t = \arccos \left( \frac{(vT)^2 + R^2 - r^2}{2vTR} \right)
\]

(37)

where \(\frac{R-r}{T} \leq v \leq \frac{R+r}{T}\). Note that, if \(v < \frac{R-r}{T}\) the UE will not suffer from a HOF, even if the entry angle \(\theta_e=0\). Likewise, if \(v > \frac{R+r}{T}\), there will not be a HOF either.

As explained previously, the RLF occurs after the UE has moved over the HOF region for more than a particular time, usually 1s. If we define this maximum time as \(T_R\), the UE will suffer from a RLF if the distance covered within the HOF circle is above \(vT\) while a HO is not yet successfully completed. Thus, the maximum angle for which the distance covered by a UE within the HOF region is larger than \(vT\), is given by

\[
\theta_R = \arcsin \left( \frac{R^2 - \left(\frac{vT}{2R}\right)^2}{R^2} \right)
\]

(38)

where \(\frac{R-r}{T} \leq v \leq \frac{2r}{TR}\). Otherwise \(\theta_R\) is not defined and a UE will not suffer from RLF. It is worth noting that, for a given entry angle \(\theta_e = \theta_R\), the UE intersects the HOF circle at two points. Thus, if we denote the distance covered by the UE from the entry point to the first intersection as \(d_1\) and to the second intersection as \(d_2\), they may be expressed as

\[
d_1 = \sqrt{R^2 - r^2 + \left(\frac{vT}{2}\right)^2 - \left(\frac{vT}{2R}\right)^2}
\]

(39)

\[
d_2 = d_1 + vT
\]

(40)

These two points will be used later on to characterize the HO, the RLF and the PP probabilities.

As initially stated in 2.4.1, a UE causes a PP if, after completing an inbound HO (which entire process lasts \(T\)), the ToS is below the MTS, hereafter referred to as \(T_{min}\). Note that the outbound HO process is triggered when the UE trajectory intersects the circumference with radius \(S\), and it is finally completed after the outbound \((TTT+T_p)\) expires, namely \(T\). Therefore, a UE will cause a PP if, after completing a successful HO, it covers a total distance within the small cell (from the entry point to the exit point)
shorter than $v(T + T_{\min} - T')$. Hence, the maximum entry angle for which a UE will not cause a PP ($\theta_s$) is given by

$$\theta_s = \arccos \left( \frac{R^2 - S^2 + v^2(T - T_{\min} - T')^2}{2Rv(T - T_{\min} - T')} \right)$$  \hspace{1cm} (41)$$

where,

$$\sqrt{S^2 - R^2} \leq v \leq \frac{R + S}{T - T_{\min} - T'}$$  \hspace{1cm} (42)$$

It may be easily calculated that a UE trajectory with an entry angle $\theta_e = \theta_s$ will not intersect the HOF circle if

$$v < \frac{\sqrt{R^2 - r^2 + \sqrt{S^2 - r^2}}}{T - T_{\min} - T'}$$  \hspace{1cm} (43)$$

Then, if this UE trajectory does intersect the HOF circle, the distances covered from the entry point to the HOF circle intersections, denoted as $d_3$ and $d_4$, can be expressed as

$$d_3 = \delta - \sqrt{r^2 - R^2 + d^2}$$  \hspace{1cm} (44)$$

$$d_4 = \delta + \sqrt{r^2 - R^2 + d^2}$$  \hspace{1cm} (45)$$

$$\delta = \frac{R^2 - S^2 + v^2(T - T_{\min} - T')^2}{2v(T - T_{\min} - T')}$$  \hspace{1cm} (46)$$

Based on the aforementioned definitions, expressions for $P_{\text{HO}}$, $P_{\text{HOF}}$, $P_{\text{RLF}}$ and $P_{\text{PP}}$ will be derived in the sequel.

According to the definitions previously stated, a UE can only perform a successful HO if the entry angle is smaller than $\theta_i$. Otherwise, the UE’s TTT timer expires after leaving the small cell coverage area and the HO process is not completed. However, and despite having $\theta_e < \theta_i$, the HO could not be completed successfully due to either a HOF (i.e. $\theta_e < \theta_i$) or an RLF (i.e. $\theta_R \geq \theta_i$ and $R - r \leq v$). If we define $\theta_{\text{HO}}^v$ as the set of entry angles that, for a given $v$, result in a successful HO, the $P_{\text{HO}}$ is then given by

$$P_{\text{HO}} = \frac{2}{\pi} \int_{\theta_e \in \theta_{\text{HO}}^v} \theta_e d\theta_e$$  \hspace{1cm} (47)$$

With regard to $\theta_{\text{HO}}^v$, it will be $\theta_{\text{HO}}^v = [0, \theta_i]$ when the HO is completed before reaching the HOF circle (i.e. $v < \frac{R - r}{T}$), $\theta_{\text{HO}}^v = [\theta_t, \theta_i]$ when $T$ expires inside the HOF circle (i.e $\frac{R - r}{T} \leq v \leq \frac{R + r}{T}$) and $\theta_R$ does not exist, $\theta_{\text{HO}}^v = [\theta_t, \theta_i]$ while $\theta_R < \theta_i$ (if $\theta_R$ exists), $\theta_{\text{HO}}^v = [\theta_R, \theta_i]$ if $vT \geq d_2$ and $\theta_R$ exists, and finally $\theta_{\text{HO}}^v = [0, \theta_i]$ when $\theta_R$ does not exist and $\frac{R + r}{T} < v \leq \frac{2R}{T}$. Thus,
\[
P_{\text{HO}} = \begin{cases} 
\frac{2}{\pi} \theta_i & \text{if } 0 \leq v < \frac{R - r}{T} \\
\frac{2}{\pi} (\theta_i - \theta_t) & \text{if } \frac{R - r}{T} \leq v \leq \frac{R + r}{T} \text{ and } \exists \theta_R \\
\frac{2}{\pi} (\theta_i - \theta_t) & \text{if } \frac{R - r}{T} \leq v < \frac{d_2}{T} \text{ and } \exists \theta_R \\
\frac{2}{\pi} (\theta_i - \theta_R) & \text{if } \frac{d_2}{T} \leq v \leq \frac{2R}{T} \text{ and } \exists \theta_R \\
-\frac{2}{\pi} \theta_i & \text{if } \frac{R + r}{T} < v \leq \frac{2R}{T} \text{ and } \exists \theta_R \\
0 & \text{otherwise}
\end{cases}
\] (48)

A HOF occurs if \( T \) expires within the HOF circle. Analogously to \( P_{\text{HO}} \), the set of entry angles for which a UE suffers from a HOF, namely \( \Theta_{\text{HOF}}^v \), is defined as \( \Theta_{\text{HOF}}^v = [0, \theta_t] \) for \( \frac{R + r}{T} \leq v \leq \frac{R + r}{T} \) and \( \Theta_{\text{HOF}}^v = \emptyset \) otherwise. Therefore,

\[
P_{\text{HO}} = \begin{cases} 
\frac{2}{\pi} \theta_t & \text{if } \frac{R - r}{T} \leq v \leq \frac{R + r}{T} \\
0 & \text{otherwise}
\end{cases}
\] (49)

The probability of RLF (RLF) is different from 0 only when \( \theta_R \) exists. Thus, assuming that \( \theta_R \) exists, the RLF presents two possible situations (see Figure 21): first, if \( R - r \leq vT \leq d_1 \), there is a RLF only after a previous HOF (and consequently, the set of angles that cause a RLF is \( \Theta_{\text{HOF}}^v = [0, \theta_t] \)); second, when \( d_1 < vT \leq 2R \), only UEs with an entry angle below \( \theta_R \) suffer from RLF (i.e. \( \Theta_{\text{HOF}}^v = [0, \theta_R] \)). Therefore,

\[
P_{\text{RLF}} = \begin{cases} 
\frac{2}{\pi} \theta_t & \text{if } \frac{R - r}{T} \leq v \leq \frac{R + r}{T} \\
\frac{2}{\pi} \theta_R & \text{if } \frac{d_2}{T} \leq v \leq \frac{2R}{T} \\
0 & \text{otherwise}
\end{cases}
\] (50)

The Ping Pong probability (\( P_{\text{PP}} \)) is defined as the probability that, for a UE that has completed a successful handover to the small cell, the duration of the connection to the small cell is below the MTS, denoted as \( T_{\text{min}} \). As the final expression of the \( P_{\text{PP}} \) is complex, let us define a set of conditions based on the definitions stated at the beginning of this paragraph.
The first condition, $C_1$, defines the range of $v$ for which $\theta_R$ exists. Note that, if there is an RLF, there cannot be a PP. Condition $C_2$ guarantees the existence of $\theta_s$. If $C_2$ does not hold, either $C_3$ or $C_4$ must be true. In particular, $C_3$ means that the ToS of the UE in the small cell is long enough (due to small $v$) to achieve $P_{pp}=0$. As for $C_4$, the speed is too high and so $P_{pp}=1$. When $C_5$ is accomplished, the trajectory of a UE with $\theta_e = \theta_s$ does not intersect the HOF circle. Condition $C_6$ is particularly important, since it means that $\theta_s \geq \theta_i$. It is worth noting that, for all $\theta_i \leq \theta_s$, there is not any PP, whereas for all UEs with $\theta_i > \theta_s$ there cannot be a HO. Hence, if $C_6$ holds, the $P_{pp}=0$. Finally, $C_7$ is equivalent to $\theta_R \geq \theta_s$, whereas $C_8$ is equivalent to $\theta_t > \theta_s$.

In the following, we will denote the complement of a condition $C_i$ as $C_i'$. For instance, when $C_2$ holds, neither $C_3$ nor $C_4$ do. Therefore, $C_2' = C_3 \cup C_4$. Based on the definitions, $P_{pp}$ may be expressed as

\[ P_{pp} = \begin{cases} 0 & \text{if } C_2 \text{ or } C_6 \text{ hold} \\ 1 & \text{if } C_4 \text{ or } C_8 \text{ hold} \\ \end{cases} \]

### 2.4.3 Numerical Results

The scenario under study consists of a small cell overlaid with a macro cell, with a distance between the macro cell site and the small cell site ($D$) that ranges from 40m to 240m. UEs are spread randomly over the layout moving at 60km/h (a very high speed in an urban scenario) and heading for the small cell coverage area with a random entry angle and moving on a straight line. The transmitted power of the macro and the small cell is 46 and 20 dBm, respectively [47],[48]. Although a range of 16 possible $TTT$ values are defined in [31], only the results (both simulated and analytical) for an illustrative subset of them have been included (specifically, $TTT$ equal to 128, 256, 320 and 512 ms). The rest of the simulation parameters can be found in Table 4.

Focusing on $P_{HO}$ and $P_{HOF}$, depicted in Figure 22 and Figure 23, it is important to point out their dependency on $\theta_i$, $\theta_e$, $R$ and $r$. In particular, the inspection of (48) reveals that $P_{HO}$ grows when $\theta_i$ grows and/or $\theta_e$ falls. Thus, for a given $v$ and $T$, $\theta_i$ rises when $R$ grows (or in other words, when $D$ is increased) according to (36). Conversely, it may be observed in (37) that $\theta_i$ decreases as $R$ rises. These two factors result in the upward trend shown in Figure 22.
Table 5 Simulation Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISD for Macro cell</td>
<td>500 m</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Macro and Small Cell Frequency</td>
<td>2 GHz</td>
</tr>
<tr>
<td>Macro cell Path-Loss</td>
<td>(128.1 + 37.6 \log_{10}(\text{distance})), (distance in km)</td>
</tr>
<tr>
<td>Small cell Path-Loss</td>
<td>(140.7 + 36.7 \log_{10}(\text{distance})), (distance in km)</td>
</tr>
<tr>
<td>Macro cell transmitted power</td>
<td>46 dBm</td>
</tr>
<tr>
<td>Small cell transmitted power</td>
<td>20 dBm</td>
</tr>
<tr>
<td>HO A3 Hysteresis Margin</td>
<td>3 dB</td>
</tr>
<tr>
<td>TTT values</td>
<td>128, 160, 320, 640 ms</td>
</tr>
<tr>
<td>HO Preparation Time</td>
<td>50 ms</td>
</tr>
<tr>
<td>HO Execution Time</td>
<td>40 ms</td>
</tr>
</tbody>
</table>

\[ P_{pp} = \begin{cases} 
0 & \text{if } C_3 \\
1 & \text{if } C_4 \\
0 & \text{if } C_2 \cap C_6 \\
\frac{2}{\pi} (\theta_i - \theta_s) & \frac{P_{HO}}{P_{HO}} \text{ if } C_1 \cap C_2 \cap C_5 \cap C_6 \cap C_7 \cap C_8 \\
\frac{2}{\pi} (\theta_i - \theta_s) & \frac{P_{HO}}{P_{HO}} \text{ if } C_1 \cap C_2 \cap C_5 \cap C_6 \cap C_7 \cap C_8 \\
1 & \text{otherwise} 
\end{cases} \quad (52) \]

As for HOF probability, \(P_{HOF} \neq 0\) if \(R - \frac{r}{T} \leq v \leq R + \frac{r}{T}\). Therefore, it is tightly coupled with the size of the coverage area \(R\) and the HOF region \(r\). The numerical solution of (35) shows that in the simulated scenario the ratio \(\frac{r}{R} = \beta\) remains approximately constant and equal to 0.73 for the whole range of \(D\). Based on this, and making use of (49), the \(P_{HOF}\) will be different from 0 when

\[ \frac{vT}{1 + \beta} \leq R \leq \frac{vT}{1 - \beta} \quad (53) \]

Figure 23 displays the performance of \(P_{HOF}\) described in (53). Specifically, \(P_{HOF}\) presents an initial upward trend (commenced when \(R = \frac{vT}{1+\beta}\)) followed by a subsequent decreasing trend (that results in \(P_{HOF} = 0\) when \(R > \frac{vT}{1-\beta}\)). This also explains the peaks observed in Figure 22 for \(D \equiv 50m\) when \(TTT\) equals 320ms, and \(D \equiv 80m\) when \(TTT\) is 512ms. Initially, the \(P_{HO}\) for these \(TTT\) values has only the
contribution of $\theta_i$, since $vT < R - r$ (see the expression in (48)), thereby resulting in an abrupt increase. However, when $R = \frac{vT}{1 + \beta}$ the increase of $\theta_i$ is counteracted by $\theta_t$, causing the aforementioned peak.

Finally, Figure 24 and Figure 25 complement the analysis with the $P_{RLF}$ and the $P_{PP}$, respectively. As expected, it may be observed that $P_{RLF}=0$ as long as HOF region is not large enough to yield $vT \leq 2r$ (the necessary condition stated in (38) to have RLFs). Moreover, $T$ must be long enough so that the UE does not complete successfully a HO, while the time spent inside the HOF region is at least equal to $T_R$. Therefore, $P_{RLF} \neq 0$ for large $D$ and $T$.

With regard to $P_{PP}$, it is correlated with $P_{HO}$, since there cannot be a PP if there is not any previous HO. Figure 25 reveals that, for a given $v$, the performance in terms of PP depends on two key parameters: $D$ and $T$. In particular, when the small cell is deployed close to the macro cell site (small $D$ values) and there is a successful HO, both $R$ and $S$ are small and so the UE is prone to short ToS and the consequent PP. On the contrary, $P_{PP}$ falls for the same reason when the distance between the macro and the small cell rises. The impact of $T$ on $P_{PP}$ is in turn the opposite, since larger $T$ leads to longer HO delay, and therefore shorter ToS in the small cell (Figure 25).

Figure 22 - Figure 25 demonstrate that the appropriate selection of $TTT$, as a function of $v$ and $D$, presents important challenges to cope with the conflicting trends experienced by the different HO performance probabilities. Note that these challenges are not addressed by related works where $TTT$ remains constant for different inter-site distances. In more detail, the $TTT$ selection should start with the limitation of the maximum $P_{RLF}$ (limited to 2% according to [38]). Then, the final $TTT$ value should be selected based on the maximum acceptable $P_{HOF}$ and $P_{PP}$. It is worth noting that there is a degree of freedom in the selection of the $TTT$. This degree of freedom allows the possibility to manage the usage of the small cell. Thus, in a scenario characterized by high mobility, an increase of $P_{HO}$ of UEs moving at high speed results in a higher small cell usage. Conversely, in low mobility scenarios, $P_{HO}$ should be decreased to reduce the small cell usage (in case of overload in the small cell).

### 2.4.4 Performance results

In this work, a Handover performance analysis in terms of $P_{HO}$, $P_{HOF}$, $P_{RLF}$, and $P_{PP}$ was carried out. Specifically, we have proven the dependency of the Handover on the inter-site distance between the small cell and the overlaid macro cell center in a two-tier scenario. Furthermore, closed-form expressions for the aforementioned probabilities were derived as a function of inter-site distance and speed of the UEs. Finally, it has been shown that the appropriate $TTT$ selection, based on inter-site distance and UE's profile, is essential for maintaining the conflicting trends of different HO performance probabilities. Therefore, it should be selected in a more flexible way, on a UE basis, and adjusted to network's characteristics and objectives to enhance HO performance [50].
Figure 23: HOF Probability ($P_{\text{HOF}}$)
2.5 Work plan and Future results

2.5.1 Self-Organizing Users
As we have already seen, the data burden in cellular networks due to the high penetration of smart devices is expected to be addressed via the introduction of multi-tier networks. The utilization in a large scale of small cells will augment networks' capacity and offer the ability to serve high rates with lower cost. However, the constant densification of cellular networks with small cells will ultimately reach a limit. Structural limitations related with interference, energy consumption, spectrum availability and backhaul capacity will restrict further expansion of the network with more nodes.

The Self-Organization of the networks with the automation of procedures and shifting of the human factor to higher levels as well as the improved mobility management schemes will create breathing space for the operators in a short term. Unfortunately, in the long term new solutions should be derived in order for the networks to continue operating efficiently. Some of these issues are placed in their true dimensions in [51] where is suggested that despite major operators' views on how they can serve the expected increase in mobile traffic, their estimations are not feasible since it is not possible for cellular operators neither to ten fold their available spectrum nor to achieve a spectral efficiency gain of 10-24 times.

Furthermore, restrictions on the infrastructure cost caused by installing more and more small cells are outlined in [52] along with the subsequent increase in energy consumption. It is argued that there is a direct relation between both the energy cost and total access network cost with the deployment's density. This relation is also highlighted with the realization that even the idle power of the base stations will be a significant limiting factor. Finally, high interference, especially in hotspots, could counteract in HetNets’ vision to serve the mobile traffic optimally; in a short geographically area with over existence of cellular nodes operating simultaneously, the interference created could not only limit the spectral efficiency but also to reduce it in some cases.

2.5.2 Future objectives

Our future research will focus on the enhancement of short-range communications (i.e D2D). We will exploit their ability to re-utilize networks' resources offering higher capacity, energy reduction, and better spectrum efficiency among others. In this framework we will study the concept of self-organization on a UE basis. In that sense the acquired knowledge concerning the necessity of SO algorithms in order to address effectively variations in the performance will be helpful. Moreover, the need for coordination to accomplish the parallel operation of more than one functions and the importance of MM in future HetNets will be our guidelines in this effort. We target on the creation of clusters made of UEs with the ability to act as small cells. In these clusters the communication will be feasible not only between a pair of UEs as in D2D but between and in parallel with any UE that belongs to this cluster. Furthermore this communication, without the involvement of BSs, will not depend on whether these UEs are collocated thus removing any geographically restrictions. Moreover, UEs in such clusters could share their backhaul connections with other members of the cluster thus upgrading the overall quality of communication and increasing the capacity of the network. Because of that, UEs with low received signal quality (e.g. cell edge UEs) could benefit from this communication inside the self-organized cluster. The combination of the above could extend networks' abilities and offer flexibility as well as further improvements in system's capacity and energy consumption. Remaining in the same area, we envision that these clusters created by collaborating UEs could be organized in a distributed manner by one of the UEs. These UEs which will have the role of the so-called Cluster Head will be responsible among others for the mobility management of the UEs that wish to enter or leave the cluster. In other words the cluster head should choose in a smart way the members of the clusters in respect to minimization of distance between UEs as well as guaranteeing backhaul connection of good quality. We strongly believe that the idea of self-organized clusters created by collaborating UEs will be the next big step in future cellular networks' evolution. In addition, it looks a promising direction for future networks to address efficiently the ever increasing mobile traffic since the continuously deployment of small cells will be obstructed by architectural limitations and unmanageable costs.
• The concept of self-organization on the UE side is expected to boost networks' capacity. We will study how cooperative transmission of UEs inside the clusters along with efficient interference management, in the uplink, can result in capacity gains. We will propose schemes that manage to organize clusters in a way that each cluster has the best possible connection with the network. In that sense, we will investigate which are the best criteria to choose a cluster Head, meaning a UE that will be responsible for controlling each one of the clusters, coordinating the introduction or the exit of one UE in the cluster, managing the resources and collaborating with neighboring cluster Heads.

• We will study the QoS for cell edge UEs which is connected with the reliability of the network at these areas. We will investigate ways for improving the coverage and the reliability in a way that the QoS of cell edge UEs is enhanced. We will propose algorithms for efficient cooperative transmission inside these clusters and resource allocation schemes that will prioritise cell edge UE that usually suffer from degraded service. We strongly believe that the exploitation of the SO idea along with the use of short range communications inside these clusters of UEs, will result in continuously optimization of networks’ operation and offer better service to UEs that belong to these clusters and are located at cell edges.

• The role of QoE is becoming increasingly important since the trend is towards real time services that demand high rates. However, at this moment, most proposals are currently QoS oriented. Therefore, we aim also at self-organizing algorithms in the framework of D2D users for enhancing the QoE.
3. Algorithms for SONs and cognition for Traffic Offloading with small-cells

3.1 State of the Art on traffic offloading with small-cells

3.1.1 Financial aspects

The boost of the data traffic in cellular networks during the past few years forebodes future needs for additional capacity. Based on these predictions, both industry and academia have put the focus on solutions to increase the efficiency of the networks, thereby compensating such expected traffic growth. The densification of the Radio Access Network (RAN) with the deployment of small cells, and the resulting increase of the spectral efficiency, has emerged as one of the most promising solutions to address the problem.

Nevertheless, and despite its potential, network densification incurs important initial capital expenditure (CAPEX) as well as additional operational expenditure (OPEX) for mobile operators. These financial restrictions, that in principle could limit the actual deployment of dense networks, open up new business opportunities for third parties. In this described scenario the third parties, owners of small cells networks, will play the role of Small cell Service Providers (SSPs). Their infrastructure will be located at hotspots, either indoors or outdoors, where the Mobile Service Providers (MSPs) have increased capacity needs, and will create a heterogeneous network composite of several parties (i.e., MSPs and SSPs) to deliver pervasive broadband cellular access.

The aforementioned networks, despite coping with the current and future capacity needs, in turn pose new challenges. The arisen challenges are twofold: on the one hand the technological feasibility must be analyzed; on the other hand, an economic incentive must be generated to all the involved parties to assure the long-term sustainability.

Relevant studies have addressed this problem from a game theoretical approach, and most of them by modelling it with different auction schemes [53]-[62].

A multiple reserve price based auction mechanism that considers imprecise valuations, named EasyBid, is proposed in [53]. In order to describe the quality of a truthful auction, the authors introduce the notions of Perceived Valuation, Partial Truthfullness, and Imprecision Loss. Heuristic algorithms are used in order to conduct truthful auctions, considering that the sellers have knowledge of their perceived valuations, which could differ from the true ones. The proposed scheme shows a performance similar to an optimal Vickrey-Clark-Groves (VCG) auction, when assuming precise valuations.

An iterative double auction mechanism, which is managed by an independent broker is proposed in [54]. This mechanism is named IDA and is used for a market where MNOs lease third-party deployed WiFi or femtocell APs to offload their traffic. The broker is responsible for guaranteeing truthful bidding, thereby maximizing the market’s efficiency while maintaining the profit. The mechanism
consists of two stages during each iteration. In the first stage, the MNOs bid for every AP for their offloading needs, and the APs bid for every MNO for their serving costs. In the second stage, the broker decides on the distribution of the APs’ resources based on the previous bids. IDA satisfies the desirable economic properties, and maximizes the welfare of the market.

A reserve price auction mechanism for spectrum sharing in Cognitive Radio (CR) networks is proposed in [55]. This CR network consists of a primary spectrum owner (PO), and multiple primary (PU) and secondary (SU) users. The SUs can buy spectrum bands from the PO through the proposed auction scheme. Based on the assumption of channels with different qualities, the SUs can have different valuations for each channel. Moreover, the PO sets a reserve price for each channel according to its quality. The auction mechanism manages to allocate efficiently the available spectrum, while maximizing the SUs’ valuations.

A VCG auction-based incentive framework for accessing selfish femtocells is studied in [56]. In it, a multi-unit reverse auction is used for a single macrocell user scenario, whereas a double auction scheme is designed for multiple macrocell users. In both scenarios the auctions have been designed based on the VCG auction, and are conducted by the MSP. The results prove that the framework guarantees each agent’s individual rationality and truthfulness, hence preventing market manipulations, and finally it improves the network’s performance.

An auction-based incentive framework for leasing on-demand resources is presented in [57]. The MSP conducts reverse auctions with the third parties, generating the incentive for them to sell their capacity. The third parties are owners of WiFi APs. The area under study is divided in macrocell-sectors that contain regions, where the WiFi APs are located. By assuming that a third party may own APs in more than one region, inter-region competition is introduced among them in addition to the existing intra-region competition. The proposed algorithm, named iDEAL, takes into account variable spatial coverage of resources, as well as dynamic traffic demands. Furthermore, it improves the system’s efficiency along with the social welfare, guarantees truthful bidding, and guards against collusion. Finally, by avoiding to use the cellular resources for serving the peak traffic in a region, the algorithm lowers significantly the cost for the MSP.

A reverse auction for the purchase of ACcess Permission (ACP) from multiple hybrid access femtocell owners is presented in [58]. A Wireless Service Provider (WSP) conducts the auction, which also allows him to lease part of the each femtocell’s available capacity, until fulfilling the requiring demand. The authors assume the possibility of overlapping femtocells, which introduces competition among the latters’ owners. The problem is firstly studied with a VCG-based reverse auction, which guarantees truthful bidding, and maximizes the social welfare. Due to its high computational complexity, the author propose a suboptimal auction scheme, which is based on a two-stage greedy algorithm. The proposed scheme performs similarly to the optimal VCG mechanism in terms of social welfare and reduction of the WSP’s costs, while reduces substantially the computational complexity.

A marketplace where an MNO can lease unused capacity from residential users in order to serve the excessive traffic of its mobile users is presented in [59]. The residential users, owners of indoor APs, act as the sellers of the marketplace. The MNO needs to lease the proper APs, which will serve the high mobile traffic, as well as maximize the market’s social welfare. This allocation problem is formulated as a combinatorial reverse auction, which is proven to be truthful, resilient against market manipulation, and minimizing the MNO’s leasing costs. Due to the auction’s high computational complexity (NP-Hard), the authors propose a greedy algorithm that solves efficiently the allocation problem in polynomial time, for a variety of real-size large network scenarios.

The problem of an MNO incentivizing femtocell owners to lease their unused spectrum resources is addressed in [60]. The MNO bases its selection on the preferences that it has for each femtocell. These preferences are a function of the femtocell’s spectral efficiency, and the price the femtocell owners set for their spectrum resources. Furthermore, the MNO’s subscriber are described by the type of their traffic (voice, video, data), and their demand which changes dynamically with time. The problem of maximizing the system’s efficiency is solved with a VCG-based reverse auction mechanism, which is
described by individual rationality and truthfulness. However, due to the high computational complexity of this mechanism, the authors further propose a sub-optimal two-stage greedy algorithm. The greedy algorithm not only preserves the characteristics of individual rationality and truthfulness, but also performs similar to the optimal one.

An incentive framework for motivating users to offload their traffic according to their delay tolerance is studied in [61]. In this work, the MNO needs to incentivize users with high delay tolerance and large offloading potential in order to increase the system’s efficiency, and minimize the total incentive cost. In order to solve this problem, the author proposes a reverse auction-based incentive mechanism, named Win-Coupon. In this auction scheme, the MNO is the buyer. It offers coupons to its subscribers, asking in return their delayed service through traffic offloading. The users place bids, where they inform the MNO about the maximum delay they can tolerate, along with the corresponding coupon in exchange for the delay. Then, the allocation problem is solved optimally by the MNO, while taking into account the offloading potential of the users, along with historical data, such as the user’s mobility patterns. Finally, the efficiency of the mechanism is proved by extensive trace-driven simulations.

An open market where a mobile operator can lease capacity from access points of independent third parties is proposed in [62]. In this work, the offloading problem is formulated as a combinatorial reverse auction. Its objective is the selection of the cheapest access points, and the offloading of the maximum possible volume of traffic. Regarding offloading, the operator can cover partially its demand from each access point, due to their limited resources. In order to guarantee both individual rationality and incentive compatibility, an innovative payment rule based on the VCG scheme is proposed. The allocation problem is solved optimally, however due to its high computational complexity (NP-hard), the authors further propose three greedy algorithms. The simulation results of realistic network scenarios show that the greedy algorithms have a performance similar to the optimal one, solving the allocation problem in polynomial time, while maintaining the truthfulness property.

Apart from the auction schemes, many works in the literature use different approaches from game theory. The works in [63] and [64] make use of Nash bargaining. The data offloading problem for a MNO monopoly is studied with the use of Nash Bargaining in [63]. In this scenario, the MNO bargains with third party access point owners (APOs) for offloading its traffic. This one-to-many bargaining problem is solved with Nash Bargaining Theory. The authors compare the sequential and concurrent bargaining in terms of social welfare. Moreover, they study the effects of the two bargaining methods on both the MNO and the APOs. Finally, the authors analyze how the APOs’ grouping affects the bargaining results.

The cooperation of a mobile operator (MO) with a fixed-line operator (FO), in order to provide femtocell service to indoor users is studied in [64]. The MO demands from FO an indoor femtocell service with certain QoS requirements. In order to achieve an agreement, the MO offers to share its profits with FO, provided that the latter deploys a low cost infrastructure. When they reach an agreement for this service, they operate as virtual integrated operators. The scenario is formulated with a sequential game and Nash bargaining. The sequential game describes the relation between the MO and the users. First, the MO sets the price for the femtocell service, and then the users select their spectrum requirements based on the price. Nash bargaining is then used between the two operators in order to share fairly the profit.

The operators will cooperate only if their collaboration results in more revenue for both of them. Thus, the evaluation of the agreement is based on the comparison of the net profits in the cooperation and non-cooperation cases. The results verify that both operators achieve higher revenues, along with a substantial increase in the spectral efficiency.

The traffic offloading through a third party, with WiFi or femtocell access points (APs), is considered in [65]. It is assumed that the offloading can occur between one macrocell BS and multiple APs, and vice versa. A market based solution is introduced, where economic incentives are given for offloading traffic. In order to find a balance between the amount of traffic that should be offloaded and its financial
compensation, the authors propose a two-stage multi-leader multi-follower game, called data offloading game (DOFF), where BSs act as leaders and APs as followers.

In the first stage of the game, all the BS offer prices to the APs in their coverage areas, whereas in the next stage, the APs decide on the volume of traffic they will offload based on the proposed prices. DOFF reaches an equilibrium that exists between two extremes, which are classic market outcomes; the Market Balance and the Monopoly Outcome.

A scenario where a macrocell service provider (MSP) needs to incentivize SSPs to accept its excessive traffic with a monetary compensation is studied in [66]. Although the scenario is similar, the situation differs, since the service providers’ subscribers (whether MSPs’ subscribers or SSPs’ subscribers) also play a major role in it. They can dynamically select their service provider based on the provided QoS and the cost, which depend on the pricing and an open access ratio. The latter is defined as the ratio of the SSPs’ resources used by the macro users and their entire volume of resources in a small cell.

A hierarchical dynamic game framework is proposed for modelling the interactive decision model. In the lower level, an evolutionary game is designed in order to describe the users’ dynamic service selection. In the upper level, the service providers’ decisions regarding their pricing strategy and the open access ratio are formulated as a Stackelberg differential game. In it, the MSPs are the leaders, whereas the SSPs are the followers. Their decisions are based on the outcome of the evolutionary game, that is, the status of their users.

Another case of traffic offloading to privately owned femtocells and the respective economic framework is described in [67]. The femtocells operate in a hybrid access mode in order to accommodate public users. In order to provide incentives to the femtocell owners for serving its traffic, the MNO proposes profit sharing.

The economic framework is modelled by a two-stage sequential game is modelled, named Femtocell Service Selection Game (FSSG). In the first stage, the operator chooses the ratio of the revenue distribution to femtocell owners in order to maximize its own benefit. In the second stage, the femtocell owners decide the proportion of the spectrum available for the public users, while ensuring their own subscribers’ service requirements. At the same time, all the subscribers select from which tier they will be served. It is shown that the concept of sharing motivates the femtocell owners to share their resources, while providing the operator with more served users, and hence increased profits.

A scenario where the femtocell operator initiates the network sharing arrangements is studied in [68]. Macro UEs can connect to the femtocell, which operates in a hybrid access mode. The macro UEs have to rent the power resource of the femtocell, which is assumed to be limited. If the entire femtocell’s power resource is used while one of the femto UEs enters the femtocell’s coverage area, the femtocell selects one macro UE to act as a relay for the additional femto UEs. When this occurs, the relaying macro UE is served by the femtocell free of charge. The problem of optimal power allocation and price setting is solved with the use of a Stackelberg game, where the macro UEs are the leaders, whereas the femtocell is the follower. The macro UEs set the demand, and then the femtocell decides the price. This game’s equilibrium provides the optimal solution, which results in the improvement of the throughput of both macro and femto UEs.

In the hybrid offloading policies, the offloading decision is shared between the user and the network. This provides a tradeoff in terms of decision complexity and robustness between a decision totally controlled by the network and user respectively. The network manages the overall policy framework by deriving the policies considering the operator strategies and network conditions. These policies are periodically sent to users which contribute to achieving the network objectives by making an optimum offloading decision and maximizing their own performance by combining network information with their own measurements and service requirements. This hybrid approach results in enhancing the overall system performance through optimized offloading decision while at the same time the user experience is improved as well. In order to realize this hybrid approach, we propose an architecture and mechanism which is based on principles of self-management and autonomic networking.
A policy based offloading model, which is based on a cost function approach is studied in [69]. The policies under study are user centric, network centric and hybrid. In the user centric approach, a network selection decision is made by the user after performing inter-system measurements, while taking into account the UE capabilities and service requirements. Similarly in the network centric approach, the network is responsible for the above procedure.

The lack of information in the previous cases is tackled by the hybrid policy approach, where the offloading decision is made cooperatively between the network and the user. For this case, a mechanism based on autonomic networking principles is proposed. This mechanism chooses the policies dynamically, according to the traffic load and the operator strategies. The authors use realistic traffic loads for the evaluation of the mechanism, which shows that the hybrid approach outperforms the other two, in terms of offloading efficiency and blocking ratio.

Similar to [58], an ACcess Permission (ACP) transaction framework is proposed in [70]. This framework enables an MSP to buy ACP from multiple (geographically overlapping) femtocell service providers (FSPs). In [70], the scenario’s problem is the dynamic change in the MSP’s demand both in time and space. This leads to incomplete information, which in turn impedes the FSPs from choosing an appropriate strategy. This problem is solved with the use of an adaptive strategy updating algorithm, which is based on an online learning process. A theoretical proof is given that the payoff gap between the proposed algorithm and the compared optimal strategy is bounded. The simulation results show that the FSPs’ profit depends on the learning speed of the proposed algorithm, as well as the information on the MSP’s load that is available to them.

A model on user service adoption is described in [71]. The users need to choose to be served between a basic wireless technology and bundle, which consists of the basic technology and a supplementary one. The supplementary technology is used by the service provider in order to offload traffic from the basic technology’s network. The choice of the technology depends on the technology’s intrinsic qualities, the throughput degradation due to congestion, as well as the service’s access prices offered by the ISP.

It is assumed that the ISP runs a monopoly. The model of the dynamics of user adoption is formulated according to the user’s payoff for each of the service provided. The choice of the service is based on the user’s valuation of each option. The choices evolve over time according to fluctuations in the congestion levels. It is shown that the user adoption reaches a unique, stable equilibrium, and that the population condition in the ISP’s network is a critical factor for the adoption behaviour equilibrium.

The interaction between an operator and its subscribers for adopting delayed WiFi offloading is studied in [72]. This is modelled as a two-stage sequential game, where the operator (leader) sets the pricing, while the users (followers) act as price-takers. Regarding the users, the authors consider their willingness to pay, their traffic demand, their tolerance in delay, and their movement patterns, which decides the probability of WiFi coverage. As for the operator, four different pricing schemes are examined: flat, volume, two-tier and congestion pricing. The simulation results give incentives to both operator and users for adopting delayed WiFi offloading, due to the financial benefits it offers.

The problem of mobile data traffic offloading through a third-party WiFi AP from the cellular operator’s perspective is studied in [73]. The authors assume a usage-based pricing model, and formulate the offloading problem as a utility maximization one. The authors examine the impact of successive interference cancelation (SIC) on the operator’s utility. They consider three cases regarding its availability: at both the operator’s base station (BS) and the AP, at neither the BS nor the AP, and only at the BS. The three scenarios are solved (near) optimally with a centralized approach, and prove that SIC benefits the operator regarding the maximization of its utility. In order to solve the problem in a computationally efficient way, the authors propose a threshold-based distributed data offloading scheme, for the case when SIC is available at both the BS and the AP. The simulation results confirm that this algorithm performs similarly to the centralized scheme, when a proper threshold is chosen.

A queuing analytical model for delayed mobile data offloading is proposed in [74]. The analysis is based on 2D Markov chains, while the derived expressions for the performance metrics are functions of
the maximum delay of the flows, which is defined as a deadline. A number of scenarios with realistic WiFi network availability statistics were studied. Based on these scenarios, the authors solved optimization problems, which involved the system delay, monetary and energy costs, by manipulating properly the maximum deadlines. The solution of the optimization problems resulted in the derivation of the achievable trade-off regions as a function of the network parameters.

A WiFi offloading model that brings mobile Internet Protocol (IP) integration to a core network with policy and charging control (PCC) is proposed in [75]. The authors design an analytical model that evaluates the offloading performance, and examines the probability of missing a deadline (maximum time period for an offloading session), as well as the volume of the cellular resources saved by offloading. The provided numerical results show that a substantial volume of cellular resources is saved, along with the satisfaction of the deadline assurance.

A framework that models the relationship between a MNO and a small-scale operator (SSO) in the context of WiFi offloading is presented in [76]. In this framework, the SSO sets the offloading price per subscriber, and the MNO determines the amount of users to be offloaded in order to minimize its costs. In this scenario, the problem is to find the SSO price that maximizes the two entities’ utilities. The proposed model is applied in a real WiFi deployment, providing insight regarding the financial aspects (costs and benefits) of such collaborations.

A cellular-to-mesh (C2M) data offloading scheme for LTE-A users to WiFi mesh networks is presented in [77]. These C2M networks can be either community networks or networks managed cooperatively by residential users, and can be leased by MNOs for offloading their traffic. The authors introduce an analytical framework for deciding the users to be offloaded, having as criterion the energy cost their demands incur to the LTE-A base stations. A routing policy is designed, so that the mesh network serves the offloaded traffic with the minimum possible cost. In order to incentivize the mesh network users for participating in the joint task, the authors use the Shapley value profit sharing rule on the reimbursement they receive from the MNO. The results show that the application of this C2M scheme reduces significantly the eNB’s power consumption, and provides satisfying profits to the mesh users.

A traffic offloading to small base stations (SBS) scheme, where the SBSs cache and deliver content to a MNO’s subscribers is propose in [78]. The interaction of the SBS owners with the MNO is formulated with a two-stage non-cooperative Stackelberg game, whose objective is the minimization of the MNO’s costs. In the first stage the MNO decides the prices paid to the SBS owners, along with the caching and routing policy. In the second stage the SBS owners determine separately the optimal strategy for storage and bandwidth allocation, based on the MNO’s reimbursement prices from the first stage. This minimization cost problem is NP-hard, therefore the authors further propose an iterative Prime-Dual algorithm. The simulation results present the efficiency of the offloading scheme, as well as the fast convergence of the proposed algorithm to the optimal solution.

### 3.2 A novel learning mechanism for traffic offloading with Small Cell as a Service

#### 3.2.1 Scenarios, methodology and performance analysis

In contrast to the SoA, this work analyzes the traffic offloading problem under the Small Cell as a Service (SCaaS) approach [79]. In these scenarios, the third party, hereafter referred to as Small Cell Operator (SCO), is the owner of small cell infrastructure, composed of a cluster of co-channel operating small cells and the backhaul network to transport data from the small cells to the Internet or to the MSP core network. However, unlike the MSPs, the SCO does not have licensed spectrum and so it must be provided by the MSP. In particular, there is an economic transaction between the SCO and each MSP that determines the maximum MSP traffic that can be offloaded to the SCO, the spectrum sub-bands licensed to the MSP that can be used by the SCO, and finally the price that the MSP pays for the small cells infrastructure usage. The problem is modeled with an auction scheme in which the SCO is the auctioneer. In this context, and in order to adapt each actor’s decisions to the traffic variations, we
propose a novel learning mechanism that consists of a traffic forecasting method, a reinforcement learning algorithm and an adapting search range scheme for improving the MSPs strategies. In the following, we provide the basic motivation and challenges of the SCaaS approach.

3.2.1.1 SCaaS

In general, the main objective of the MSPs is to meet the traffic demand of their customers with the already deployed network, thereby maximizing the economic gain. However, the ever increasing traffic demand makes it imperative to densify the RAN to boost the spectral efficiency. In this context, SCaaS emerges as one of the major cost-effective alternatives to conduct the aforementioned RAN densification, since it is conceived as a single infrastructure (usually deployed by a third party known as SCO) to service a set of MSPs. As first introduced in the above, the MSPs have the exclusive right to use the licensed spectrum, and therefore the SCO can only use such spectrum after mutual Service Level Agreement (SLA) with the corresponding MSPs. Due to the need for a regulatory framework to support these SLAs, a huge effort has been done to provide consistent spectrum regulation policies in different regions, e.g, the spectrum manager leasing arrangement or the de facto leasing arrangement [80].

Focusing on the financial aspects of the SCs network deployment, the initial investment required to cover high traffic zones all over the coverage area of the MSPs may be unaffordable for a single actor. The SCaaS approach allows MSPs to distribute such high CAPEX among several SCOs (i.e. different SCOs in different locations), but in turn this CAPEX reduction is translated into an increase of the MSPs OPEX via the price to use the SCs network. Therefore, CAPEX and OPEX of the different stakeholders are tightly coupled.

It is precisely due to the link between MSPs and SCOs financial structure that, for a given SCaaS deployment, it is necessary to strike a balance between the improvement of the MSPs spectral efficiency (by offloading traffic through the SCO network) and the costs associated to the SCO network usage. In turn, the SCO revenues must be high enough to recoup the deployment investment, though low enough to stimulate offloading strategies of the MSPs. This situation becomes even more complex in competitive environments, where several MSPs compete in an auction for the usage of the limited SCO resources. Therefore, SCaaS pose new techno-economic challenges to reach long-lasting equilibria that foster the emergence of third parties willing to invest in dense network deployments, and the efficient use of the deployed capacity.

The rest of this section is organized as follows; firstly we describe the system model. Then, we present the auction scheme, the involved parties’ objectives and the learning mechanism. Finally, we provide numerical results and analysis.

3.2.1.2 System Model

The scenario under study is composed of $N$ eNBs that belong to $N$ different MSPs, namely $MSP_i$ with $i = 1 \ldots N$, and a single SCO cluster. In turn, the SCO cluster consists of a set of $N_{sc}$ outdoor Home eNBs (HeNBs) deployed over a high traffic area (i.e., a hotspot) and a backhaul network with a maximum capacity $C_{BH}$.

Each $MSP_i$ may be characterized by its spectral efficiency, $SE_i$, and its licensed bandwidth denoted by $B_i$. As for the offered load of the $MSP_i$ (referred to as $L_i$), it may be divided into two components: the offered load generated within the hotspot ($L_{hi}$) and the offered load generated elsewhere ($L_{ni}$), i.e., $L_i = L_{hi} + L_{ni}$. Likewise, the SCO may also be characterized by the spectral efficiency, $SE_{sc}$, and the maximum bandwidth supported by the small cells according to the hardware limitations set by the respective technology, denoted by $B_{sc}$.

Load variations, both in time and space, are one of the key aspects of the stated problem. Thus, during time periods where the offered load in the hotspot (i.e., $L_{hi}$) is low, the need for the additional capacity provided by the SCO declines. Conversely, high hotspot loads result in a raising interest for the usage of the SCO infrastructure. In this sense, the day is divided into a set of $T$ equal timeframes $t_n$, with $n = 1 \ldots T$, during which the load can be regarded as a constant. It is nonetheless worth noting that,
even though the load in a timeframe \( t_n \) is not necessarily the same every day, it is not independent from the load in the \( n \)th timeframe of the previous and subsequent days; in other words, the load tends to follow a day pattern.

At the beginning of each timeframe, an auction is started by the SCO. Each \( MSP_i \) bids an amount of money \( b_i \) to use the SCO infrastructure. After receiving the bids of all the MSPs, and based on them, the SCO distributes its capacity among the bidders. Thus, the \( MSP_i \) will be allowed to offload a maximum load through the SCO equal to

\[
L_{si}^{sc} = N_{si} x_i B_i S_{E_{sc}},
\]

where \( x_i \in [0,1] \) is the part of the bandwidth provided by the \( MSP_i \) and used by the SCO. As defined before, the bandwidth allocated to the SCO cannot exceed \( B_{sc} \).

\[
B_{sc} \geq \sum_{i=1}^{N} x_i B_i \quad (3.1)
\]

In turn, the offloaded traffic must be lower than the SCO backhaul capacity

\[
C_{BH} \geq \sum_{i=1}^{N} L_{si}^{sc} \quad (3.2)
\]

### 3.2.1.3 The Auction

The distribution of the existing resources, i.e., the available spectrum and the cost of using the SCO infrastructure, are the response to two main objectives: firstly, the maximization of the throughput (the capacity objective), and secondly the optimization of the profits (the economic objectives of both the MSPs and the SCO). In this scenario the auction is the mechanism by which the interaction of these competing objectives results in incentives for all the involved parties.

#### 3.2.1.3.1 The Auction Mechanism

In any auction there are three aspects that must be defined: the auctioneer, the bidders and the good to be auctioned. In the scenario under study, the capacity of the SCO is the auctioned good, the SCO plays the role of the auctioneer and the MSPs are the bidders. Let us define the total load served by the SC cluster as

\[
L_{sc} = \sum_{i} L_{si}^{sc} \quad \text{(in Mbps)}
\]

If the distribution of the available SC capacity is based on proportional fairness [81], the maximum load that a given \( MSP_i \) can offload to the SCO is given by

\[
L_{si}^{sc} = \frac{b_i}{b_i + \sum_{j \neq i} b_j} L_{sc} \quad (3.3)
\]

Accordingly, the maximum SC capacity gained by \( MSP_i \) after bidding \( b_i \) is subject to the bids of the rest of the MSPs and to the minimum profit demanded by the SCO. Given that neither the own load is known by \( MSP_i \) (since the auction is played at the beginning of the timeframe) nor the bids and the load of the rest of the MSPs are public (i.e., \( b_j \) and \( L_j, \forall j \neq i \) are unknown), \( MSP_i \) must estimate \( L_{hi}, L_{ni} \) and \( \sum_{j \neq i} b_j \) to properly select \( b_i \). Hence, the more accurate the estimations are, the better the result of the auction.

In the following, a detailed analysis of the objectives of each party is presented to fully describe the auction. After this, a learning mechanism is proposed to enhance the bidding of the MSPs.

#### 3.2.1.3.2 The SCO’s objective

The SCO aims to maximize its profit by leasing high volumes of SC capacity at increased prices. The profit is defined as the difference of the MSPs’ bids minus its expenses that are divided into the load and fixed costs, denoted as \( C L_{sc} \) and \( C F_{sc} \) respectively. Hence, the profit can be written as

\[
\]

\footnote{Let us henceforth denote as \( b_{j \neq i} \) the sum of the opponents’ bids.}
\[ P_{sc} = \sum_{j} b_j - CL_{sc} - CF_{sc} \] (3.4)

To ensure its objective, the SCO introduces a minimum profit, which is described as a percentage of its total costs

\[ P_{sc}^{\text{min}} = (z - 1)(CL_{sc} + F_{sc}), z > 1 \] (3.5)

If \( P_{sc} < P_{sc}^{\text{min}} \), the SCO will not lease SC capacity to the MSPs. Regarding the load costs, they are given by convex functions that have been widely used in the literature [54], [57], [64], [79] for describing network congestion costs as well as the effect of subscriber churn. For low traffic, the decreased costs describe the already invested CAPEX for operating the network. Conversely, during peak traffic periods the cost increases rapidly, depicting the economic consequences of congestion. The load costs of the SCO are expressed as

\[ CL_{sc} = \frac{a_{sc}L_{sc}^2}{C_{BH} + d_{sc} - L_{sc}}, L_{sc} \in [0, C_{BH}] \] (3.6)

where factor \( a_{sc} \) defines the rate with which the cost increases (in €/Mbps), whereas \( d_{i} \) (in Mbps) moves the asymptotic discontinuity of the cost function to \( L_{sc} = C_{BH} + d_{sc} \). Note that when the SCO network operates at its maximum capacity (\( C_{BH} \)), load costs are high but not infinite. Therefore, \( d_{sc} > 0 \).

Finally, using (3.3), (3.5) and (3.6) in (3.4), the reserve price for offloading \( L_{sc}^{mc} \) traffic, denoted by \( b_{im}^{\text{min}} \), can be written as

\[ b_{im}^{\text{min}} = z \frac{a_{sc}L_{sc}^{sc}}{C_{BH} + d_{sc} - L_{sc}} + zCF_{sc} \frac{L_{sc}^{sc}}{L_{sc}} \] (3.7)

Note that the reserve price is a theoretical concept only valid when all MSPs bid in a complete information environment. In such a case, all MSPs are interested in bidding the reserve price to minimize the cost for offloading.

3.2.1.3.3 The MSPs’ objectives

The MSPs’ objectives can be characterized by their financial and network aspects. Firstly, the MSPs aim to maximize their profit, which is defined as the revenue from serving traffic minus the expenses that are categorized into load (\( CL_{i} \)) and fixed (\( CF_{i} \)) costs, in addition to the bid \( b_{i} \) for leasing SC capacity. Therefore, the profit can be written as

\[ P_{i}(x_{i}) = R_{i}(L_{i}^{mc} + L_{i}^{sc}) - CL_{i} - b_{i} - CF_{i} \] (3.8)

where \( L_{i}^{mc} \) denotes the load served by the macrocell and \( R_{i} \) the revenue per Mbps.

Secondly, they need to guarantee their subscribers’ QoS, which can be accomplished by maximizing the throughput. When offloading occurs, the total load served by \( MSP_{i} \), \( L_{i}^{T} \) is given by

\[ L_{i}^{T} = L_{i}^{mc} + L_{i}^{sc} = (1 - x_{i})B_{i}SE_{i} + N_{sc}x_{i}B_{i}SE_{sc} \] (3.9)
Since the throughput is a function of the transferred bandwidth, it is important to define the maximum and minimum limits of \( x_i \) (\( x_i^{\text{max}} \) and \( x_i^{\text{min}} \) respectively) for satisfying \( L_i^T = L_i \). When \( L_i > C_i^2 \) is assumed, they can be expressed as follows

\[
\begin{align*}
  x_i^{\text{min}} &= \frac{L_i - B_i SE_i}{B_i (N_{sc} SE_{sc} - SE_i)} \\
  x_i^{\text{max}} &= \frac{L_{hi}}{N_{sc} B_i SE_{sc}}
\end{align*}
\] (3.10)

Similar to the SCO, MSP\(_i\)'s load cost is also described by a convex function of \( L_{imc} \)

\[
CL_i = \frac{a_i (L_{imc})^2}{C_i + d_i - L_{imc}}, L_{imc} \in [0, C_i]
\] (3.11)

The values of \( a_i \) and \( d_i \) characterize the load cost function in (3.11). The costs can also be expressed in terms of Marginal Cost (MC) [64],[71] which is defined as the change in total cost for producing a unit of the offered good (\( L_{imc} \))

\[
MC = \frac{dCL_i}{dL_{imc}} = \frac{a_i L_{imc}^2 (2C_i + 2d_i - L_{imc})}{(C_i + d_i - L_{imc})^2} \left[ \frac{€}{Mbps} \right]
\] (3.12)

In order to highlight the dependencies of the characteristic cost parameters (i.e. \( a_i \) and \( d_i \)), the MC function is evaluated for \( L_{imc} = C_i \),

\[
MC(C_i) = \frac{a_i C_i (C_i + 2d_i)}{d_i^2} \Rightarrow a_i = \frac{MC(C_i) d_i^2}{C_i^2 + 2 C_i d_i}
\] (3.13)

Applying the same analysis to the SCO, and assuming that both the MSPs and the SCO share the same congestion costs (\( CL_{sc} (C_{BH}) = CL_i (C_i) \)), \( a_{sc} \) can be written as

\[
a_{sc} = \frac{d_{sc} CL_i (C_i)}{C_{BH}^2}
\] (3.14)

Solving the system of (3.13) and (3.14) leads to the values of \( a_{sc} \) and \( d_{sc} \).

3.2.1.3.4 Learning mechanism

The necessity of using a learning mechanism was pointed out in subsection 3.2.1.3.1. However, in order to present the mechanism, it is essential to describe first the MSPs’ decision making policy. This refers to the way the MSPs decide on the traffic to be offloaded to the SCO \( L_i^{sc} \), as well as the corresponding bid at each auction. It was assumed that the MSPs have information about the SCO’s distributing and charging mechanism, that is, equations (3.3) and (3.7). By rearranging (3.7) with respect to \( L_{sc} \), the following expression is obtained

\[
z a_{sc} L_{sc}^{2} + (b_{j \neq i} - z C F_{sc}) L_{sc} - (C_{BH} + d_{sc})(b_{j \neq i} - z C F_{sc})
\] (3.15)

Solving the system of (3.3) and (3.15) results in MSP\(_i\)'s offloaded traffic \( L_i^{sc} \). The solution of this new equation provides a function \( b_{i} (L_{i}^{sc} | b_{j \neq i}) \) that gives the winning bid for a demand \( L_i^{sc} \), given the sum of \( ^2 \) Let us denote as \( C_i = B_i SE_i \) the MSP\(_i\)'s macrocell sector capacity, when the former does not offload traffic at the SC cluster.

Security: Public
the opponents’ strategies $b_{j|\neq i}$. This information would enable $MSP_i$ to decide how much money it needs to pay for maximizing $L^*_i$ and $P_i$. However, such information is not available. As a result, the MSPs need to choose their strategies based on estimates of $b_{j|\neq i}$. Hence, $MSP_i$’s bidding strategy selection results from maximizing $P_i$ given an estimate $\hat{b}_{j|\neq i}$, while setting a constraint on the minimum throughput. This is formulated as

Maximize:

$$P_i(x_i|\hat{b}_{j|\neq i}) = R_i(L_i^{mc} + L_i^{c}) - CL_i - b_i(x_i|\hat{b}_{j|\neq i})$$

Subject to:

$$x_i^{\text{min}} \leq x_i \leq x_i^{\text{max}}$$

Note that the limits imposed to $x_i$ in (3.10) ensure the availability of resources, both in the hotspot and elsewhere, to serve the offered load of $MSP_i$. Therefore, the MSPs prioritize their throughput over their profits. The decision making could be described as $MSP_i$’s best response to $\hat{b}_{j|\neq i}$.

We are now in a position to introduce the learning mechanism, which was designed in order to assist the MSPs’ bid selection strategy. As stated above, it depends on $MSP_i$’s demand and its opponents’ bids, indicating the importance of accurately estimating $L_i$ and $b_{j|\neq i}$ at each timeframe. Consequently, our proposed learning mechanism consists of three components. A forecasting method for predicting $L_i$, and a reinforcement learning algorithm along with an adaptive search range scheme for estimating $b_{j|\neq i}$.

1) Traffic Forecasting: In order to forecast the MSPs’ traffic loads $L_i$ and $L_{hi}$, the well-known and computationally efficient Holt-Winters (HW) method [82] was selected. Also known as Triple Exponential Smoothing, it takes into account the level, trend and seasonal changes in the observed dataset. There are two HW models according to the type of the seasonality\(^3\), known as multiplicative and additive seasonal models. The former refers to a proportional change in the values of the time series from season to season, whereas the latter refers to a particular absolute change. In our case, due to the behavior of the traffic pattern that is described by random, small, seasonal variations, the following equations for the multiplicative model were used

$$S_t = a \frac{L_t}{I_{t-T}} + (1 - a) (S_{t-1} + v_{t-1})$$

$$v_t = \beta (S_t - S_{t-1}) + (1 - \beta) v_{t-1}$$

$$I_t = \gamma \frac{L_t}{S_t} + (1 - \gamma) I_{t-T}$$

$$\hat{L}_{t+m} = (S_t + m v_t) I_{t-T+m}$$

where $L_t$ denotes the observation of the offered load during timeframe $t$, and $S_t$ a (smoothed) estimate of the level, that is, a local average of the dataset. Parameter $v_t$ is an estimate of the linear trend (slope) of the time series, whereas $I_t$ denotes the seasonal component, in other words, the expectation for a specific timeframe based on its past season values. Parameter $\hat{L}_{t+m}$ denotes the forecast at $m$ timeframes ahead, $t$ the current timeframe, and $T$ the number of daily timeframes (season length). In our

\(^3\) Seasonality can be described as the periodic repetition in time-series data
scenario \( m = 1 \), since the estimate of the next timeframe was needed. The three smoothing factors \( \alpha, \beta, \gamma \in [0, 1] \) show the dependence on the past values of the time series. Lastly, the method’s parameters are initialized as

\[
S_t = \frac{1}{T} \sum_{t=1}^{T} L_t, v_t = 0, I_t = \frac{L_t}{S_t}, t = 1 \ldots T
\]

For the initialization of the seasonal factors, a general rule suggests the use of at least two seasons of historical data.

2) **Learning Algorithm**: For a given continuous set of aggregate bid values \( A_{b_{j 
eq i}} = [b_{j 
eq i}^{\min}, b_{j 
eq i}^{\max}] \) that contains \( b_{j 
eq i} \), the learning algorithm should be able to estimate \( b_{j 
eq i} \). Nevertheless, using a continuous set could prove inefficient in our scenario, since the convergence requires thousands of iterations. Considering that, \( A_{b_{j 
eq i}} \) was discretized and the reinforcement learning algorithm Exp3 \([83]\) was used. By doing so, the action selection was set up as a non-stochastic multi-armed bandit problem, where each \( b_{j 
eq i} \in A_{b_{j 
eq i}} \) was regarded as one arm.

We propose Algorithm 4.1 to estimate the opponent’s auction strategy, which is a modification of [83]. For each timeframe, an action set \( A_{b_{j 
eq i}} \) is defined and discretized to \( K \) values. When the algorithm is initialized in day \( d = d_0 \), each value is given an equal probability \( p_t = 1/K \). Then, an action is drawn randomly based on the probability distribution, and the corresponding \( b_i \) is placed according to (3.16) and (3.17). A reward system is used for updating \( p_t \). The rewards \( r_t \) are based on comparisons between the expected and the real results of the auction. As stated before, the MSPs know the SCO’s distributing and charging mechanism. Hence, they can use their estimates \( \bar{L}_i \) and \( \bar{b}_{j 
eq i} \) to emulate the auction and its results. Since the actual bids are sealed, the gained SC capacity \( L_i^{SC} \) was chosen for the reward system. Particularly, the relative difference of the expected and actual value, \( \Delta L_i^{SC} \) was mapped to a reward \( r_t \), denoting that for \( \Delta L_i^{SC} = 0 \) the reward is \( r_t = 1 \), whereas for \( |\Delta L_i^{SC}| \geq 0.5 \) it is \( r_t = 0 \).

**Algorithm 1 Estimation of opponent’s auction strategy**

1. Initialization: \( d = d_0 \)
   For one timeframe \( t \)
   Define \( A_{b_{j 
eq i}} = [b_{j 
eq i}^{\min}, b_{j 
eq i}^{\max}] \) and discretize it to \( K \) values \( b_{j 
eq i}^k, k = 1 \ldots K \).
   Set probabilities: \( p_t(d_0, b_{j 
eq i}^k) = 1/K, \forall b_{j 
eq i}^k \in A_{b_{j 
eq i}} \)
2. Draw \( b_{j 
eq i}^k \) randomly based on the probability distribution and place corresponding \( b_i \) according to (16) and (17).
3. Calculate relative difference
   \[ \Delta L_i^{\hat{r}} = \frac{(E[L_i^{\hat{r}}] - L_i^{\hat{r}})}{L_i^{\hat{r}}} \]
4. Map \( \Delta L_i^{\hat{r}} \) to a reward \( r_t(d_n, b_{j 
eq i}^k) \in [0, 1] \)
   \[ r_t = -2|\Delta L_i^{\hat{r}}| + 1 \]
   If \( \mu_{\Delta L}(d_n-1 \ldots d_n) < 0.15 \) AND \( \sigma_{\Delta L}(d_n-1 \ldots d_n) < 0.2 \)
   \[ r_t = -\frac{1}{\mu_{\Delta L}} |\Delta L_i^{\hat{r}}| + 1 \]
5. For \( l = 1, 2, \ldots, K \) calculate the expected reward
   \[ r_t(d_n, b_{j 
eq i}^k) = r_t(d_n, b_{j 
eq i}^k) / p_t(d_n, b_{j 
eq i}^k) \] if \( l = k \)
   \[ r_t(d_n, b_{j 
eq i}^k) = 0 \] otherwise,
   update and normalize probability distribution
   \[ p_t(d_{n+1}, b_{j 
eq i}^k) = p_t(d_{n+1}, b_{j 
eq i}^k) \exp(r_t(d_n, b_{j 
eq i}^k)) / K \]
   \[ p_t(d_{n+1}, b_{j 
eq i}^k) = p_t(d_{n+1}, b_{j 
eq i}^k) / \sum_{k} p_t(d_{n+1}, b_{j 
eq i}^k) \]
Furthermore, another rule was included for improving already increased rewards. If the average $\Delta L_i^{sc}$ of the current and the last 4 days $\mu_{dL}(d_{n-4} \ldots d_n)$, as well as the standard deviation $\sigma_{dL}(d_{n-4} \ldots d_n)$ are bellow a certain value, $r_i = -\frac{\Delta L_i^{sc}}{\mu_{dL}} + 1$. A small $\mu_{dL}$ denotes good estimations, whereas a small $\sigma_{dL}$ ascertains that these estimates do not reside within a large range of values, but they are close enough to $\mu_{dL}$. The values in the condition were chosen after extensive simulations ($\mu_{dL}(d_{n-4} \ldots d_n) < 0.15$ AND $\sigma_{dL}(d_{n-4} \ldots d_n) < 0.2$). If it holds, the played action will receive a low reward, forcing the algorithm to explore neighboring values. Subsequently, the expected reward $\hat{r}_i$ is calculated according to step 5, and all the probabilities are updated. Finally, $\eta \in (0, 1]$ denotes the learning speed; the higher its value, the faster the probabilities increase.

3) Search Range Scheme: In order for Algorithm 4.1 to work, $A_{b_j=1}$ must contain $b_{j=1}$. Hence, if the actual aggregate bid does not lie in the initial estimate of the action set, the algorithm will not be able to predict it. Then again, even if the initial estimate is correct, the algorithm will not be able to follow any changes of $b_{j=1}$ outside the predefined range. Therefore, a variable parameter space scheme was used [84]. The aim of this scheme is to center the probability distribution obtained by the learning algorithm around a certain value by expanding and reducing its action set.

In this work, it was utilized as an extension of the learning algorithm, so that $b_{j=1} \in A_{b_j=1}$. This scheme introduces parameters that determine the speed of the changes in $A_{b_j=1}$. These are the expansion and reduction rate, expressed as $E_r$ and $R_r$ respectively. Updating the action set rapidly or slowly might result in getting farther from the actual bid or not converging fast enough, respectively. The choice of expanding or reducing the range depends on conditions applied to the probability distribution. The parameters used in the conditions for updating $A_{b_j=1}$ are the minimum ($b_{j=1}^{min}$) and maximum ($b_{j=1}^{max}$) action values, as well as the 25th ($b_{j=1}^{25}$) and the 75th ($b_{j=1}^{75}$) percentile. The latter are compared with other percentiles that are defined by the expansion ($E_r$) and reduction ($R_r$) coefficients. In order for the range $A_{b_j=1}$ to expand or reduce, one of the following conditions [C 1]-[C 5] must be satisfied. The conditions for expansion [C 1]-[C 5] and reduction [C 6] are defined as:

[C 1] $b_{j=1}^{25} < b_{j=1}^{25-E_r}$ AND $b_{j=1}^{75} > b_{j=1}^{75+E_r}$

[C 2] $b_{j=1}^{25} > \left( b_{j=1}^{max} + b_{j=1}^{min} \right)/2$

[C 3] $b_{j=1}^{25} = b_{j=1}^{75}$

[C 4] $p_t(d, b_{j=1}) = 1/K \; \forall b_{j=1} \in A_{b_j=1}$

[C 5] $\mu_{dL}(d_{n-4} \ldots d_n) > 0.2$ AND $\Delta L_i^{sc}(d_n) < 0.15$

[C 6] $b_{j=1}^{25} > b_{j=1}^{25+R_c}$ AND $b_{j=1}^{75} < b_{j=1}^{75-R_c}$

Conditions [C 1] and [C 6] are from the initial scheme in [84]. The remaining conditions were introduced for specific cases that could not be handled by [C 1] and [C 6]. Condition [C 2] is used when the distribution’s mode is at $b_{j=1}^{max}$. The same applies to [C 3], but for the lower endpoint $b_{j=1}^{min}$. Condition [C 4] is used when the algorithm is in a stalemate due to the actual bid being far from $A_{b_j=1}$, so that $r_i = 0$. Finally, [C 5] is examined only when none of the rest is satisfied. It is applied when the mechanism is in a stalemate and does not place $A_{b_j=1}$ closer to the real bid. The new $A_{b_j=1}$ endpoints are given by:
Note that, when $A_{b_{j \neq i}}$ is expanded or reduced, the range for which the probability distribution is defined vary. Therefore, when the range is reduced the probability distribution at the new points is set at the interpolated values, whereas for expansion these values are obtained through extrapolation.

Additional rules were introduced for changing dynamically the learning, expansion and reduction rates in order to improve the learning mechanism. The rationale is that the MSPs should use high rates when $\mu_{\Delta L}$ is increased, in order to detect faster the opponent’s bid. On the other hand, when $\mu_{\Delta L}$ is below a certain value, they should reduce the rates for not deviating from a good prediction. The conditions are:

\begin{align*}
\text{[C 7]} & \quad \mu_{\Delta L}(d_{n-4} \ldots d_n) > 0.2 \\
\text{[C 8]} & \quad \mu_{\Delta L}(d_{n-4} \ldots d_n) < 0.15 \text{ AND } \sigma_{\Delta L}(d_{n-4} \ldots d_n) < 0.15
\end{align*}

The new rate values are:

\begin{equation}
\begin{align*}
\eta & \leftarrow +\eta + 0.1 \\
E_r & \leftarrow 1.2E_r, \text{ for [C 7]} \\
R_r & \leftarrow 1.2R_r \\
\eta & \leftarrow +\eta - 0.1 \\
E_r & \leftarrow 0.5E_r, \text{ for [C 8]} \\
R_r & \leftarrow 0.9R_r
\end{align*}
\end{equation}

3.2.1.4 Performance Evaluation

In this section, we evaluate our mechanism’s performance by presenting the accuracy of its estimations, and its comparison with the ideal case of Complete Information (CI), where actual values are used instead of estimations. A scenario with two MSPs that share the same parameter values (given in Table 6 and Table 7) was implemented. Our work is focused on 3GPP’s LTE-A networks, therefore, our values were chosen based on relevant studies [85]. Regarding the offered load, a traffic shape based on a bimodal distribution was used. Hence, a 24-hour pattern was devised with two peaks at $t_{12}$ and $t_{16}$. The actual offered load of each MSP is generated from the pattern. Thus, in every timeframe, the offered load is a random variable centered at the load pattern value.

The mechanism’s steady state (where MSPs converge) was compared with an ideal case that assumes CI for both MSPs, in other words, they know their opponent’s network and financial parameters. This case was regarded as a multi-objective optimization problem, where the goal was the maximization of the profits (objective functions) of all the parties. Since no preferences were assumed, the Objective Sum Method was used [86] which maximizes the sum of the three profits, providing always a Pareto optimal solution. The now single objective optimization problem was formulated as

Maximize: $P(x_1, x_2) = P_1 + P_2 + P_{sc}$
Subject to: (3.2) and (3.17),

where the first constraint denotes the SC cluster capacity limitation, whereas the second one guarantees that $L^{x_i}_i = L_i$. It should be noted that since the bids are subtracted, function $P(x_1, x_2)$ only depends on the MSPs’ revenue and the participating parties’ costs. As a result, maximizing $P$ is equivalent to minimizing the parties’ costs, which can be achieved by serving $L_i$ with the minimum use of transferred spectrum $x_i$.

Before presenting the results, it should be noted that the hotspot traffic $L^{h}_i$ ranges between 20 – 60% of $L_i$. When offloading occurs, the macrocell’s capacity $C_i$ is reduced to $L^{mc}_i$, while the available SC capacity is $L^{sc}_i$, since a part $x_i$ of MSP’s bandwidth is used at the SC cluster. During the low traffic timeframes ($L_i < C_i$), the MSPs offload their traffic to increase their profits. This happens because offloading small volumes of traffic is cheaper than serving it with the macrocell. Hence, our focus is mainly on the peak traffic hours (i.e., $t_{11-17}$), where $L_i > C_i$ and $L^{h}_i$ ranges between 50–60% of $L_i$.

Table 6: MSP-SCO Network and Financial parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth</td>
<td>20 MHz</td>
</tr>
<tr>
<td>Sector Spectral Efficiency</td>
<td>1.7 bps/Hz</td>
</tr>
<tr>
<td>Transferred resources</td>
<td>$x_i$ [0.0, 0.28]</td>
</tr>
<tr>
<td>Revenue per timeframe</td>
<td>$R_i$ 3.375 €/Mbp</td>
</tr>
<tr>
<td>Marginal Cost</td>
<td>$MC(C_i)$ 9.45 €/Mbp [8]</td>
</tr>
<tr>
<td>Cost shaping factor</td>
<td>$a_x$ 0.405 €/Mbp</td>
</tr>
<tr>
<td>Cost shaping constant</td>
<td>$d_x$ 10 Mbp</td>
</tr>
<tr>
<td>Fixed Costs per sector</td>
<td>$C_{F1}$ 1.72 €/h</td>
</tr>
</tbody>
</table>

Table 7: Learning Mechanism parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constants</td>
<td>$\alpha$ 0.2, $\beta$ 0.1, $\gamma$ 0.1</td>
</tr>
<tr>
<td>Learning Algorithm</td>
<td></td>
</tr>
<tr>
<td>Granularity of $A_{3g}$</td>
<td>$K$ 3</td>
</tr>
<tr>
<td>Learning Speed</td>
<td>$\eta$ 0.3</td>
</tr>
<tr>
<td>Adaptive Search Range</td>
<td></td>
</tr>
<tr>
<td>Percentage expansion coefficient</td>
<td>$E_{\alpha}$ 1</td>
</tr>
<tr>
<td>Percentage reduction coefficient</td>
<td>$E_{\beta}$ 17</td>
</tr>
<tr>
<td>Expansion Rate</td>
<td>$E_{r}$ 0.3</td>
</tr>
<tr>
<td>Reduction Rate</td>
<td>$R_{r}$ 0.5</td>
</tr>
</tbody>
</table>

As it can be observed in Figure 26, the mechanism is able to properly estimate $b_2$, since the error is below 5% in most of the timeframes. The bad estimate in $t_{18}$ ($\Delta b_2 \approx 10.5\%$) does not hinder the mechanism from performing well, since the bid and the estimation error are quite small ($b_1 \approx 1.4$ €, see Figure 28) to affect significantly the profit and the auction result, respectively. The same applies to $t_{10}$. 
Figure 27 compares our proposal and the CI case with regard to the throughput and unserved traffic, while in Figure 28 they are compared in terms of profit and bid. The aforementioned good estimates provide a performance almost identical to the ideal case in comparison. The latter is marginally superior with 1.2% maximum difference for the throughput and 9.3% for the profit. Both of them are observed in $t_{12}$, where the offered load has its peak. At this timeframe, the MSPs’ increased needs for offloading cannot be covered by the SC cluster. Therefore, a small volume of unserved traffic is observed. The small differences in the throughput are due to the estimation errors of $b_2$ that are seen in Figure 26. The corresponding $b_1$ is not sufficient for winning the auction, which results in offloading less traffic.

The difference in the profit can be explained by the difference observed in $b_1$ (Figure 28). In the CI case the MSPs lease the necessary $L_{i}^{c}$ by bidding the reserve price $b_{i}^{\text{min}}$. On the other hand, our proposal is applied in a competitive environment, where due to the lack of information it has the tendency of increasing the MSPs’ bids. This behavior is justified by the conditions used for expanding $A_{b_{j
eq i}}$ in the search range scheme, in particular [C 1], [C 2] and [C 4]. This is rather intensified in $t_{12}$, where the MSPs’ demands cannot be realized concurrently, leading to a daily increase of their bids. Therefore, the bids become higher than the ones in the CI case, reducing the profits.

### 3.3 Network and financial aspects within the SCaaS approach

#### 3.3.1 Scenarios, methodology and performance analysis

This work is an extension of our previous work in [87]. Our goal is to address the interaction of the capacity needs and the economic constraints of the stakeholders, as we intend to gain insight into techno-economic implications of the SCaaS paradigm.

#### 3.3.1.1 Analysis of spectrum allocations

For this new work we retained the system model, which was used in [87]. However, since the SCaaS approach requires the transfer$^4$ of licensed bandwidth from the MSP to the SCO, the way in which the transfer is conducted will impact both the system performance and the monetary transactions.

---

4 As detailed in Section 3.2.1.1, the transfer of bandwidth is defined as a Service Level Agreement (SLA) between the legal licensee (the MSP) and the lessee (the SCO) by which the latter is allowed to use part of the spectrum of the former.
Specifically, the deployment of Heterogeneous Networks (HetNets) presents two different approaches regarding the SCs tier spectrum band allocation: the dedicated spectrum deployment and the co-channel deployment. The former (dedicated spectrum), which was studied in [87], is characterized by the orthogonal use of spectrum in the SCO and in the MSP. Thus, an $MSP_i$ that transfers a band $x_iB_i$ to the SCO only uses the non-transferred spectrum band (i.e. $(1 - x_i)B_i$). Conversely, in the latter (co-channel deployment) the spectrum is partially reused by the two tiers, and $MSP_i$ makes use of the whole band $B_i$ regardless of the transferred spectrum to the SCO, $x_iB_i$. The two approaches present pros and cons. The co-channel deployment increases the spectrum availability in the MSP after the transfer procedure at the expense of a spectral efficiency reduction (due to the interference increase). Instead, the dedicated spectrum approach leads to higher spectral efficiency though reducing the availability of spectrum in the MSP.

![Figure 29: System Model and Spectrum Allocation for the two channel deployments](image)

The system model and the spectrum allocation of the two approaches are presented in Figure 29. As it may be observed, the spectral efficiency and the available bandwidth for each player depends on how the spectrum is allocated. Thus, whereas the dedicated spectrum allocation splits the available bandwidth and takes advantage of the spectral efficiencies both in the MSP and in the SCO, the co-channel allocation increases the spectrum availability by dropping the spectral efficiency of the shared band. In the sequel, the superindexes D and C differentiate the dedicated spectrum deployment and the co-channel deployment parameters, respectively, when the impact of both alternatives on the auction mechanism is examined.

Regarding the total load served by $MSP_i$, $L_i^{CT} = L_i^{mc} + L_i^{c}$ is given by

$$L_i^{DT} = (1 - x_i^D)SE_i^D B_i + N_{sc}x_i^D SE_{sc}^D B_i \quad (3.25)$$

for the dedicated spectrum deployment, with

$$\begin{align*}
x_i^{D\min} &= \frac{L_i - B_i SE_i^D}{B_i (N_{sc} SE_{sc}^D - SE_i^D)} \\
x_i^{D\max} &= \frac{L_i B_i SE_i^D}{N_{sc} SE_{sc}^D} \\
\text{s.t. } L_i &> C_i \quad (3.26)
\end{align*}$$

And

$$L_i^{CT} = [x_i^c SE_i^c + (1 - x_i^c)SE_i^D] B_i \quad (3.27)$$

$$+ N_{sc}x_i^c SE_{sc}^c B_i$$
for the co-channel deployment. In both cases the MSP\textsubscript{i}’s macrocell sector capacity, $C_i = B_i SE_i^D$, is defined as the maximum load that can be served by MSP\textsubscript{i} without offloading traffic when the whole bandwidth is $B_i$ and the spectral efficiency is $SE_i^D$.

The two deployments’ differences in throughput can be explained through expressions (3.25) and (3.27), along with Figure 29. As it is summarized in the table of Figure 29, the deployments differ in terms of spectral efficiency and frequency reuse. For $L_i^C$, the expression is the same in both cases. The differentiating factors are the spectral efficiency and the necessary co-channel SEs as a function of the dedicated spectrum SEs. Specifically, we define $x_i^C B_i$ for offloading the same $L_i^C$. For the dedicated spectrum deployment, the macrocell’s capacity reduces linearly with $x_i^D$. Therefore, $MSP_i$ should be cautious with the volume of transferred $x_i^D B_i$ in order to be able to serve traffic outside of the SC cluster. On the contrary, the co-channel deployment’s capability of utilizing $x_i^C B_i$ outside of the SC cluster enables $MSP_i$ to use two subbands ($x_i^F B_i$,$(1-x_i^F)B_i$) with their corresponding spectral efficiencies ($SE_i^C, SE_i^D$), according to the demands that $L_i$ imposes.

Another difference in this work regarding the MSPs is that two types of small cells are considered, based on the study in [85]. These will be henceforth referred as type “T” and “E”. The former refer to street level small cells that connect to a Network Termination Point (NTP). The latter extend wirelessly from the type “T” to other type “E” small cells. By defining these two types of small cells, the fixed costs for the SCO can be expressed as

$$CF_{SC}(N_T, N_E) = CF_{SC}^T N_T + CF_{SC}^E N_E$$

where $NT$ and $NE$ stand for the number of type “T” and “E” small cells respectively, whereas $CF_{SC}^T$ and $CF_{SC}^E$ denote the corresponding fixed costs per small cell.

### 3.3.1.2 Sensitivity to spectral efficiency

As it can be inferred from the previous sections, spectral efficiency is one of the key parameters that determines the capacity of the MSP as well as the offloading capacity. In turn, the spectral efficiency values depend a great deal on aspects such as the number of SCs, their location or the bandwidth allocation, thereby impacting on the MSP’s served traffic. This impact occurs in a twofold manner: first, by limiting or increasing the achievable maximum throughput, and secondly by stimulating or discouraging the competition among MSPs for the SCO capacity (thereby increasing or reducing the MSPs’ costs). In order to gain insight into the dependencies between the throughput and the set of spectral efficiencies for the different spectrum allocations (i.e. $SE_i^D, SE_i^C, SE_{SC}^D$ and $SE_{SC}^C$), let us define the co-channel SEs as a function of the dedicated spectrum SEs. Specifically, we define $SE_i^C = \delta_i SE_i^D$ and $SE_{SC}^C = \delta_{SC} SE_{SC}^D$, where $\delta, \delta_{SC} \in (0,1)$. In turn, if we define $\Delta L_i^T = L_i^{CT} - L_i^{DT}$, and by using (3.25) and (3.27), we obtain

$$\Delta L_i^T = \left[(x_i^D - (1-\delta_i)x_i^F)SE_i^D + SE_{SC}^C (\delta_{SC} x_i^C - x_i^D)\right] B_i$$

If $\Delta L_i^T > 0$, the co-channel approach leads to a higher amount of traffic served by $MSP_i$; conversely, if $\Delta L_i^T < 0$, the dedicated spectrum approach outperforms the co-channel approach. Based on this, the preference for each approach within the SCaaS framework depends on parameters such as
\( SE^D_i, SE^D_{sc}, \delta_i, \delta_{sc} \) or \( N_{sc} \). The same effect can be observed if we assume that the traffic offloaded to the SCO by \( MSP_i \), i.e. \( L^sc_i \), is the same in both cases. In that situation, the best approach is determined by the amount of traffic served outside the hotspot, namely \( L^mc_i \). Therefore, the co-channel allocation is preferable if \( L^{mc}_i > L^{mc}_{sc} \), (and vice versa for the dedicated spectrum).

\[
SE^D_i \left[ x^C_i \delta_i + (1 - x^C_i) \right] B_i > SE^D_i \left(1 - x^D_i\right) B_i \tag{3.31}
\]

\[
(1 - \delta_i)x^C_i < x^D_i \tag{3.32}
\]

### 3.3.1.3 Conducting the auction

#### 3.3.1.3.1 Complete Information

In this section, the auction is regarded as an optimization problem. By solving it, the auction results will maximize the participants’ profits as well as their network performance. Furthermore, it will give insight into the MSPs’ auction behavior for different levels of competition, and SC deployments.

For the solution of the optimization problem the Complete Information (CI) case is introduced. Regarding our scenario, it refers to an ideal environment where the MSPs know their future traffic, and hold all the information regarding their opponents’ network and financial parameters, as well as the SCO’s distributing and charging mechanism (equations (3.3) and (3.7)). Since the MSPs and the SCO are economic entities, it is of utmost importance to them that traffic offloading is profitable. Based on that, the MSPs set their decision making policy, whose objective is the maximization of profits and throughput. Hence, the optimization problem’s solution lies in the SC capacity distribution and the corresponding payments \( (L^sc_i, b_i) \) that maximize \( P_{sc}, P_i \) and \( L^T_i \), all of them detailed in Sections 3.2.1.3 and 3.3.1.1.

In our scenario, all the stakeholders (i.e., MSPs and the SCO) must have the incentive to take part in the auction, without preference for any of the participants. Therefore, a method with no articulation of preferences was used for solving the problem. It is the “Objective Sum Method” (OSM) [86], which sums the objective functions (profits), converting the multi-objective optimization problem to a single-objective one. Moreover, the OSM’s result is always a Pareto optimal solution. Using OSM, the problem can be formulated as

Maximize:

\[
P(x_1, \ldots, x_N) = \sum_{i=1}^{N} P_i + P_{sc} \tag{3.33}
\]

Subject to:

\[
x_i^{min} \leq x_i \leq x_i^{max} \tag{3.34}
\]

\[
C_{BH} \geq \sum_{i=1}^{N} L^sc_i \tag{3.35}
\]

Note that the objective function \( P(x_1, \ldots, x_N) \) is independent of the MSPs’ bids. This means that \( P \) can be maximized by increasing \( L^i \), while reducing \( CL_i \) and \( CL_{sc} \). The limits imposed to \( x_i \) in (3.34) ensure the availability of resources, both in the hotspot and elsewhere, to serve the offered load of \( MSP_i \). However, the constraint in (3.35) sets an upper limit in \( L_{sc} \). This means that for increased \( L_i \) values, and in turn
requirements for offloading, the MSPs’ demands for SC capacity cannot be satisfied concurrently. This results in a competition among the MSPs, whose level is defined by the set \( \{ L_1, \ldots, L_N \} \). Consequently, for high \( L_i \) values the cost of offloading increases rapidly due to SCO’s congestion costs, affecting significantly \( P_i \). Under such circumstances, the MSPs’ decisions are affected due to a trade-off that appears between \( P_i \) and \( L_i^T \).

### 3.3.1.4 Performance evaluation

The scenario used for the numerical analysis in this section consists of \( N = 2 \) MSPs and a SCO with a single SC cluster in the MSPs’ overlapping macrocell coverage areas. In order to increase their profits and throughput, the MSPs participate in the auctions conducted by the SCO. The parameter values used in our simulations are listed in Table 8, which due to our focus on 3GPP’s LTE-A networks were chosen from Small Cell Forum studies [85]. It should be noted that for all the SC deployments in our simulations, the fixed costs are calculated by (3.29) for \( N_T = 1 \) and \( N_E = N_{sc} - 1 \).

The spectral efficiency values for the different deployments have been obtained through a custom-made MATLAB-based simulator. A HetNet composed of three-sectored macrocells and overlaid SC clusters have been simulated according to the guidelines provided in [20]–[22]. The results for the two approaches (co-channel deployment and dedicated spectrum deployment) are presented in Table 9.

<table>
<thead>
<tr>
<th>Table 8: MSP-SCO Network Financial Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSP</strong></td>
</tr>
<tr>
<td>Bandwidth</td>
</tr>
<tr>
<td>Sector Spectral Efficiency</td>
</tr>
<tr>
<td>Revenue per timeframe</td>
</tr>
<tr>
<td>Marginal Cost</td>
</tr>
<tr>
<td>Cost shaping factor</td>
</tr>
<tr>
<td>Cost shaping constant</td>
</tr>
<tr>
<td>Fixed Costs per sector</td>
</tr>
<tr>
<td>Hotspot offered load</td>
</tr>
<tr>
<td>SCO</td>
</tr>
<tr>
<td>Number of small cells in cluster</td>
</tr>
<tr>
<td>Max Bandwidth</td>
</tr>
<tr>
<td>Spectral Efficiency</td>
</tr>
<tr>
<td>Backhaul Capacity</td>
</tr>
<tr>
<td>Marginal Cost</td>
</tr>
<tr>
<td>Cost shaping factor</td>
</tr>
<tr>
<td>Cost shaping constant</td>
</tr>
<tr>
<td>Cost shaping factor</td>
</tr>
<tr>
<td>Fixed Costs</td>
</tr>
<tr>
<td>Profit factor</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Table 9: Average Spectral Efficiency for different small cell deployments</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_{sc} )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
</tbody>
</table>
Figure 30: Traffic served by the SC cluster versus the total offered load for MSP₁

Figure 31: MSP₁’s throughput versus the total offered load

The techno-economic aspects that SCaaS introduce are investigated in a CI environment. We will henceforth use the term use case in order to refer to a particular SC cluster deployment (i.e. the pair \(N_{sc}\) and \(C_{BH}\)). The use case under study in Figure 30 up to Figure 33 is described by \(N_{sc} = 6\) small cells and \(C_{BH} = 300\,\text{Mbps}\). Figure 30 presents MSP₁’s traffic served by the SC cluster as a function of its total offered load. The impact on the auction results caused by MSP₂’s participation is studied for particular values of its offered load, \(L₂\). The simulations were conducted for both spectrum allocations (markers were used for the dedicated spectrum deployment and lines for the co-channel deployment). As the cost function and the revenue are considered the same for MSP₁ and MSP₂, the results presented hereafter are symmetrical for both MSPs.

MSP₁ will try to offload traffic to the SCO as long as the cost of offloading is lower than the cost of serving such traffic through the eNB. This can be observed in Figure 30, where all the offered traffic in the hotspot (\(L_{h₁} = 0.6L₁\)) is offloaded for low \(L₁\) in both spectrum allocations (i.e. \(L₁^{sc} = L_{h₁}\)). Thus, for low offered loads MSP₁ offloads \(L_{h₁}\), even though it could be served by the macrocell. For higher offered loads, MSP₁ prioritizes \(L_{h₁}\) over the traffic elsewhere in order to maximize its profit, but the amount of traffic finally offloaded is limited either by the spectral efficiency of the SCs (\(SE_{sc}\)) or by the offloading cost (tightly coupled with the backhaul capacity, \(C_{BH}\)).

For any deployment, the maximum offloaded traffic is achieved when \(L₂ = 0\) (i.e. no competition). As \(L₂\) increases, and so does the competition, the SC capacity is shared among the MSPs. That is, the maximum traffic that can be offloaded by MSP₁ is reduced with the competition, mainly due to the share-out of the available backhaul and air capacity and the consequent nonlinear increase of the offloading costs. Moreover, as observed in Figure 30, the maximum achievable \(L₁^{sc}\) is then reached for lower total offered load \(L₁\) as \(L₂\) is increased. Note that for medium and high \(L₂\) values the traffic offloaded by MSP₁, \(L₁^{sc}\), is the same regardless of the deployment. This occurs because the backhaul capacity (\(C_{BH}\)) and the corresponding offloading costs are limiting the offloaded traffic. However, for low \(L₂\) (i.e. low competition) the dedicated spectrum deployment outperforms the co-channel deployment. In this case, the low level of traffic served by the SCO keeps costs low and the offloading limitation arises in the spectral efficiency. In particular, as \(\delta_{sc} = 0.46\) for \(N_{sc} = 6\) (see Table 8), the dedicated spectrum is able to increase the amount of offloaded traffic in comparison with the co-channel deployment.
Another limiting factor for the MSPs is the SCO’s congestion costs. As $C_{BH}$ is limited, the demand of SC capacity from MSPs cannot be always fulfilled concurrently. Furthermore, when their needs require the entire $C_{BH}$, they have to cover the SCO’s high congestion costs. Therefore, in order to maximize their profits, the MSPs always offload less traffic than the SC cluster’s backhaul capacity (i.e. $L_{sc} < C_{BH}$) in order to avoid these costs. This can be seen in expression (3.33), where an increase in $L_{sc}$ would increase the MSPs’ throughput, and hence their revenue. Nevertheless, the corresponding increase in $C_{L_{sc}}$ would be much higher, resulting in a lower $P(x_1, x_2)$ value. This can be clearly seen in Figure 30 for $L_2 = 0$, where the maximum achievable offloaded traffic is slightly below $C_{BH}$ (around 293 Mbps).

Also from Figure 30, it is worth inspecting the curves for both deployments with $L_2 = 245Mbps$. As it may be observed, when $L_1 \geq 245Mbps$, the traffic offloaded remains constant. Therefore, even when $L_1 \gg L_2$, $L_{sc} = C_{BH}/2$ (in fact, it is slightly below $C_{BH}/2$). After several simulations, this fact has been proven true for different deployments (i.e. different values of $N_{sc}$ and $C_{BH}$), thereby concluding that in high competition situations the SCO capacity is equally distributed among the MSPs, no matter how imbalanced their offered loads are. The rationale behind this effect must be sought in the profit and cost functions of the MSPs. Specifically, for a given MSP, e.g. $MSP_i$, the offloading of a new traffic unit will only increase the profit if the revenue minus the offloading cost (i.e. the bid, $b_i$) of the traffic unit is higher than the revenue minus the MSP serving cost ($C_{L_{sc}}$). Thus, the MSP’s bid $b_i$ is upper bounded. The same reasoning can be applied to the rest of the MSPs. If the revenues are the same for all the MSPs, and given that the SCO cost function is the same as well, their bids are also equally upper bounded, and the proportional fair based distribution of resources (see (3.3)) results in $L_{sc} = C_{BH}/N$, where $N$ is the number of MSPs. The described situation, where the SCO capacity is equally distributed among the MSPs, will be henceforth denoted by saturated competition.

Figure 31 illustrates $MSP_1$’s throughput as a function of $L_1$. It can be noticed that for this specific use case, and for $L_1 > 130Mbps$, the co-channel deployment performs better for most of the different levels of competition. The reason for that is twofold. Firstly, as it is observed in Figure 30, $MSP_1$ offloads almost the same volume of traffic with both spectrum allocations (i.e. $L_{sc} = L_{sc}^D$). Secondy, as the entire bandwidth is available in the macrocell for the co-channel and the $SE_i^{C}$ is slightly reduced ($\delta_i \approx 0.95$), then $L_{sc}^{C} \equiv C_1$. As a consequence, $L_{sc}^{C} > L_{sc}^{D}$. When $L_2 = 0$ the dedicated spectrum outperforms the co-channel approach for the same reasons described in Figure 30.

Figure 32 shows $MSP_1$’s profit versus the total offered load. It can be seen that the channel deployments’ profitability has the same trend as their network performance shown in Figure 31. This means that for every level of competition, the most efficient deployment in terms of throughput is the most profitable as well. Also, it is observed that even though the co-channel deployment achieves
considerably more throughput for high competition (see Figure 31), the corresponding difference in the profit is quite smaller.

This trend is shown clearly in Figure 33, where the ratio of MSP$_1$’s profit to its throughput (i.e. $P_1/L_1^T$ in Mbps) is plotted as a function of MSP$_2$’s total offered load, for $L_1 = 550Mbps$. It is seen that for low competition, $P_1^D/L_1^{DT} > P_1^C/L_1^{CT}$. In other words, the dedicated spectrum deployment not only outperforms the co-channel deployment in terms of throughput, but it also presents a higher profit per throughput unit. However, for low to medium $L_2$ values, $P_1^D/L_1^{DT}$ decreases faster than $P_1^C/L_1^{CT}$. This occurs because for the dedicated spectrum deployment $P_1^D$ drops faster than LDT. Since $L_{sc}^{D} \equiv C_{BH}$ for $L_1 = 550Mbps$, the bid offered by $MSP_1$ is equal to its upper bound regardless of $L_2$. Hence, as $L_2$ grows, and consequently $L_{sc}^{D}$ falls, the profit per throughput decreases until the saturated competition is reached. Conversely, in the co-channel deployment $L_{sc}^{C} \ll C_{BH}$ for $L_2 = 0$. As $L_2$ grows, $b_1$ raises, whereas $L_{sc}^{C}$ and $L_{CT}$ remain constant for $L_1 < 110Mbps$. However, the corresponding decrease in $P_1$, and hence in $P_1^C/L_1^{CT}$ is minuscule, as it is observed in Figure 33. Finally, the ratio drops until the saturation competition is reached.

![Figure 34: MSP$_1$ throughput for no competition ($L_2 = 0$)](image)

![Figure 35: MSP$_1$ throughput for saturated competition](image)

In Figure 34 and Figure 35 we present the maximum throughput achievable by $MSP_1$ as a function of the available SC backhaul capacity $C_{BH}$, for two cases: $L_2 = 0$ and saturated competition. Each curve corresponds to a SC density, whereas each subplot shows one of the spectrum allocations. In Figure 34, for both deployments, when $C_{BH}$ is low there is a linear increase in $L_1^T$. As it was mentioned for Figure 30, the MSPs do not use the entire $C_{BH}$ even if it is necessary due to the high congestion costs. Nevertheless, since $CL_{sc}$ is a function of $1/(C_{BH} + d_{sc})$, $CL_{sc}$ falls for any given $L_{sc}$ when $C_{BH}$ is increased, thereby allowing $MSP_1$ to offload more traffic without being constrained by the SCO’s costs. In short, the maximum achievable throughput is limited by the backhaul capacity. Nonetheless, the maximum achievable throughput cannot be increased by infinitely deploying additional backhaul capacity. Such maximum achievable throughput is thus upper bounded by the $SE_{sc}$ value when $C_{BH}$ is high. This is clearly shown in Figure 34, where the $C_{BH}$ limited region (low $C_{BH}$) and the $SE_{sc}$ limited region (high $C_{BH}$) may be identified. As expected, the densification of the SCO network (i.e. the increase of $N_{sc}$) can only boost the maximum achievable throughput in the $SE_{sc}$ limited region. Despite the differences existing between dedicated and co-channels deployments, the same rationale is applicable to both deployments. As for Figure 35, the same results are obtained in the saturated competition case. In this situation, however, the backhaul and air capacities are both equally shared by the two MSPs.

As observed so far, the performance in terms of throughput and profit under the described SCaaS framework is influenced by several factors, among which the backhaul capacity (with the associated congestion costs) and the SCs’ spectrum allocation come up as the most important ones. In order to gain
insight into the convenience of conducting co-channel or dedicated spectrum allocations, Figure 36 and Figure 37 explore the difference between the throughput achieved by the co-channel deployment and the throughput achieved by the dedicated spectrum deployment, defined in (3.30) as $\Delta L_i^T = L_i^{CT} - L_i^{DT}$, for cases without competition and in saturated competition respectively. According to the definition, if $\Delta L_i^T > 0$ the co-channel deployment provides higher capacity; conversely, when $\Delta L_i^T < 0$ it is convenient to transfer spectrum from the MSPs to the SCOs in a dedicated manner.

As analyzed previously, the performance of the system can be divided into the backhaul limited region and the spectral efficiency limited region. In the former, the performance of the system is similar for both deployments, though with a slightly better performance of the co-channel deployment. Contrarily, in the latter the lower spectral efficiency of the co-channel deployment results in a better performance of the dedicated allocation. These regions can be observed in Figure 36, where $L_i^T$ remains slightly positive when $C_{BH}$ is low (i.e. $C_{BH} = 100\,Mbps$ or $C_{BH} = 200\,Mbps$), since the system is limited by the backhaul capacity. In such a case, the improvement of the spectral efficiency achieved by increasing $N_{sc}$ has no significant impact on $L_i^T$. When $C_{BH}$ is high (e.g. $C_{BH}$ equal to 600 or 700 Mbps) the densification of the SCO network does improve the spectral efficiency in both deployments (more specifically, the spectral efficiency multiplied by the number of SCs), but such improvement, together with the availability of different bandwidths in each case, is more noticeable in the dedicated spectrum allocation. Therefore, $L_i^T$ decreases as $N_{sc}$ grows.

![Figure 36: Throughput comparison for no competition ($L_2 = 0$)](image1)

![Figure 37: Throughput comparison for saturated competition](image2)

Intermediate situations (in the scenario under study $C_{BH}$ from 300 to 500 Mbps) are particularly interesting. For a small number of SCs, the system is basically limited by the spectral efficiency. However, as the number of SCs grows, the traffic served by the SCO rises and the backhaul starts limiting the achievable throughput. Thus, the system transits from a spectral efficiency limited state to a backhaul limited state by simply increasing $N_{sc}$. Obviously, such transit turns up before for lower backhaul capacities. When there is competition, as shown in Figure 37, both backhaul and air capacity are shared among the MSPs and, consequently, the system transits to the backhaul limited state before.

Generally, it is observed that the MSPs aim to maximize their throughput and satisfy their subscribers. However, when traffic offloading leads to high costs, and therefore reduced profit, there is a reduction of the throughput $L_i^T$. As for the co-channel or dedicated spectrum, low competition can be better addressed with the dedicated spectrum allocation, whereas in high competition among MSPs the co-channel alternative presents better results.
3.4 Future research planning

As it has been observed in the past few years, there has been an explosive increase in the mobile traffic, a trend that will continue in the future. Furthermore, the recent advancements in technology have turned the average subscriber’s phone into a generator of large volumes of data (e.g. 4K video recording, HD online gaming etc.). Furthermore, the average user’s demand in data increases due to the use not only of his phone, but also from different devices such as smart wearables with wireless connectivity. Hence, the data traffic demand changes shape due to the parallel usage of more than one devices, which translates in the concurrent use of services. Apart from the human type communications (HTC), during the past few years a great number of devices such as sensors and security cameras have access to the internet. It is foreseen that the machine type communications (MTC) will constitute a great demand of data traffic, as they will be deployed everywhere, especially in large urban areas. Moreover, the above type of communications will not only increase the needs for network capacity. A number of new technologies will require apart from high data speeds, low latency, as well as ubiquitous coverage. However, these requirements cannot be met by the current 4G technology, LTE-A. As a result, there has been great interest in establishing the requirements of future technologies from the telecom industry side [88], in order to launch 5G networks in the not so distant future (2020).

The ESR’s current research is focused on the requirements the upcoming 5G technologies set. An urban outdoor scenario is assumed, where the traffic demand is generated by subscribers (HTC) and its requirements are set according to [88]. Also, each user is assumed to have demands for more than one service, each one having different QoS requirements. Regarding the architecture, the following are assumed. This architecture is described by a macrocell network operating in the microwave band, overlaid by a densely deployed small cell network, which works in a millimeter wave band. Moreover, since there is a difficulty in backhauling a large number of small cells in a wired number, it is assumed that they are connected wirelessly to the core network through multi-hop connections.

Figure 38: System Model

In this scenario, the ESR will study the UE association problem, that is, in which BS should each UE connect in order to maximize the aggregate subscriber utility taking into account the QoS requirements. Special focus will be given on the latency, due to the delay that is bound to appear in a multi-service scenario with the majority of the traffic volume travelling through a dense small cell network with multi-hop wireless backhaul. For this scenario, the ESR will propose a solution for the system that fulfills the QoS requirements.
4. Algorithms for SONs and cognition for decentralized network management methods, policies and algorithms

4.1 SON Introduction

The Self-Organizing Network (SON) concept is a management framework to dynamically and autonomously configure mobile networks via a set of independent SON functions that can be executed locally, centrally or in a hybrid way based on the network requirement and deployment scenario. These SON functions aim to perform dedicated network management tasks such as network configuration, optimization and failure recovery also referred to as self-healing. Various network parameters such as antenna tilts, transmit powers, and cell individual offsets etc. are automatically adjusted by a SON function for maximizing the quality of service (QoS) while minimizing operators’ capital and
operational expenses (CAPEX and OPEX). A number of sophisticated optimization techniques have been developed and studied by the researchers for example in [89], [90], rule based algorithms such as in [91], [92] or learning based techniques as discussed in [93], [94]. SON functions can operate either in online or offline mode also referred to as real time and non-real time mode. In online mode, the SON functions perform QoS optimization in real time during the network operation which means, that parameter changes and updates are implemented directly on the live network. Once this is done, the network’s response to the performed parameter modifications is measured and the future network parameter adjustment is done accordingly. In contrast to online SON functions, the offline SON functions perform QoS optimization as a simulation process using an accurate modelling environment emulating the network that needs to be optimized. Once this process is complete, the parameter setting, which seems to be optimal in the simulation, is applied to the network. In [95] authors discuss qualitative properties of on- and off-line SON solutions and examined a real-world capacity and coverage optimization (CCO) use case in order to quantitatively compare an on- and off-line SON solution. They conclude that, it is recommended to use an offline SON solution whenever the required input knowledge is available. However, online SON solutions also provide adequate alternative if accurate data on users’ location and signal strengths is missing due to for example, technical or commercial reasons.

4.2 SON Standardization Status:

The standardization efforts for SON in 3GPP have been going on since a long time. SON was introduced in 3GPP Rel-8 and has been a feature of LTE since then. This section provides a brief update on the SON standardization efforts mainly in 3GPP LTE/LTE-A.

Most of the SON features have been introduced in 3GPP releases Rel-8 and Rel-9.

For example,

Release 8:

- Automatic Neighbour Relation (ANR)
- eNB and MME self-configuration
- Physical Cell Identity (PCI) optimization

Release 9:

- Mobility Load Balancing
- Mobility Robustness optimization

The enhancements and improvements to various SON use cases have been introduced in Rel-9, Rel-10 and Rel-11 [96].

As LTE network evolved towards ubiquitous coverage, the operators’ focus shifted towards optimizing the capacity in a heterogeneous environment with different layers of cells such as macros, micro, pico and femto, having different radio access technologies (RAT) in 2G, 3G and with multiple carriers per RAT [8]. In 3GPP Rel-10, additional features to further optimize the
performance of heterogeneous networks are being introduced with the aim to further reduce OPEX. The enhancements to existing SON features considered in Rel-10 SON are as follows.

Release 10 provides enhancements to improve the reliability and mobility of load balancing in intra-LTE scenarios and inter RAT scenarios. It provides enhancements related to the mobility robustness optimization, for example to detect connection failures and provide information required for corrective actions such as in the case of handover failure. The detection of unnecessary inter-RAT handovers and reporting that events back to the source eNB were also introduced. Release 10 also provides enhancements to Inter cell Interference Coordination (ICIC) SON mechanisms for deployment scenarios with macro and femto cells. Studies in Rel-10 have shown dominant interference conditions when Non Closed Subscriber Group (CSG)/CSG users are in close proximity of femto cells. Enhancements related to Coverage and Capacity Optimizations are also included in Rel-10. In addition to this, Rel-10 also discusses proposals related to enhancements in the area of cell outage compensation, energy savings and self-healing mechanisms. For Minimization of Derive Tests (MDT), 3GPP Rel-10 defined two modes of reporting for the MDT measurements, namely, Immediate MDT and Logged MDT. A UE in connected mode is configured with Immediate MDT that implies immediate reporting. A UE in idle mode of operation is configured with Logged MDT.

Specific measurements supported for Immediate and logged MDT performances for E-UTRAN are specified in Rel-10. MDT measurement collection task are specified to be initiated in two distinct ways, first, management based MDT and second, signaling based MDT. Rel-10 has specified that the MDT data reported from UEs and the RAN may be used to monitor and detect coverage problems in the network including coverage hole, weak coverage, pilot pollution, overshoot coverage, coverage mapping and the UL coverage. 3GPP Rel-11 focuses on management aspects of ANR for UTRAN and Inter-Radio Access Technology (IRAT). The management aspects for a number of SON use cases in context with UTRAN Automatic Neighbor Relation (ANR) have been identified in 3GPP Rel-11 such as

- Intra-UTRAN ANR
- UTRAN IRAT ANR from UTRAN to GERAN
- UTRAN IRAT ANR from UTRAN to E-UTRAN

In addition to UTRAN ANR, Rel-11 has addressed management aspects for E-UTRAN IRAT ANR, such as:

- ANR from E-UTRAN to GERAN
- ANR from E-UTRAN to UTRAN
- ANR from E-UTRAN to CDMA2000

3GPP Rel-11 has also defined a combination of targets to balance the network load. These targets include
- RRC connection establishment failure rate related to load
- E-RAB setup failure rate related to load
- RRC Connection Abnormal Release Rate related to load
- E-RAB Abnormal Release Rate related to load
- Rate of failures related to handover

Some additional specific load balancing related performance measurements for use in SON includes:

- Number of failed RRC connection establishments related to load,
- Total number of attempted RRC connection establishments,
- Number of E-RAB setup failures related to load,
- Total number of attempted E-RAB setups,
- Number of abnormal RRC connection releases related to load,
- Total number of RRC connection releases, the
- Number of E-RAB abnormal releases related to load,
- Total number of E-RAB releases,
- Number of failure events related to handover and
- Total number of handover events

The handover parameter (HO) optimization is done to avoid problems such as too early handovers, too late handovers and inefficient use of network resources due to HOs. The SON algorithms can be used to perform HO parameter optimization. These optimization algorithms are not specified in the standard; however the set of HO parameters that may be adjusted by these algorithms is dictated by the choice of triggered HO measurements made by the Radio Resource Management (RRM) entity in an eNodeB. Two options for the location of the SON algorithm for HO parameter optimization have been specified in 3GPP Rel-11, namely, in the eNB(s), and in the element manager through which the parameter changes are executed in the eNBs. A HO Parameter Optimization Monitor Function has also been specified to monitor the handover parameter optimization (for example, monitoring related performance counters or alarms) and a HO Parameter Optimization Policy Control Function to be used for configuring the handover parameter optimization policies.

The HO-related performance measurements from the source and / or target eNB has also been specified in 3GPP Rel-11 which can be useful in detecting HO-related issues on the cell level such as, the number of RLF events within an interval after handover success, the number of unnecessary handovers to another RAT without Radio Link Failure (RLF) and specific performance measurements related to
handover failure (number of handover events, number of HO failures, number of too early HO failures, number of too late HO failures, number of HO failures to wrong cell, number of unnecessary HOs to another RAT). Problem scenarios are identified based on UE measurements, performance measurements, and event capture and analysis. Further improvements to the handover optimization mechanism (e.g., inter-eNB and inter-RAT transfer of RLF information using core network interfaces in areas with scattered LTE coverage) and options to evaluate the situation after a completed handover are also included as a part of 3GPP Rel-11.

The coverage and capacity optimization (CCO) in a network aims to provide optimal coverage and capacity. The 3GPP Rel-11 addresses in detail, the causes of CCO problems in the network. These causes include for example coverage hole, weak coverage, pilot pollution, overshoot coverage and DL and UL channel coverage mismatch. The performance related measurements for CCO can be collected at the source and/or target eNBs. These measurements could be useful in detection and resolution of capacity and coverage related problem at the cell level. 3GPP Rel-11 specifies inputs required for the identification of the problem scenarios such as UE measurements, performance measurements, alarms and other monitoring information, for e.g. trace data. Rel-11 has specified that a tradeoff between capacity and coverage needs to be considered. It defines the parameters such as downlink transmit power, antenna tilt and antenna azimuth that needs to be optimized to reach capacity and coverage optimization targets. It also suggests that the measurements such as Minimizing Drive Tests (MDT), HO related measurements, UE measurements etc. maybe used to detect and identify capacity and coverage related problems at the cell level. Monitoring information, for e.g. trace etc., alarms can be correlated to get an unambiguous indication of capacity and coverage problems. Rel-11 also defines logical Functions for CCO, such as, CCO Monitor Function and CCO Policy Control Function, to be used for configuring the capacity and coverage optimization policies. The option to place the centralized CCO SON algorithm in the element management or in the network management layer is defined in the Rel-11. Performance measurements related with CCO are also specified including maximum carrier transmit power and mean carrier transmit power.

A self-organized RACH optimization function is used to automatically set several parameters related to the RACH performance. The target values for access probability and access delay configured by the operators have been defined by the 3GPP. SON based RACH optimization entity is specified to be located in the eNB. Performance measurements related with RACH optimization are the distribution of RACH preambles sent and distribution of RACH access delay.

3GPP Rel-11, recognized the importance of energy saving management and suggested the network operators to minimize energy consumptions to reduce cost and protect the environment. SON algorithms in Operations, Administration and Maintenance (OAM) unit of mobile networks can contribute to energy savings by allowing the operator to set policies to minimize consumption of energy, while maintaining coverage, capacity and quality of service. The permitted impact on coverage, capacity and quality of service is determined by an operator’s policy.

Two energy saving states for a cell with respect to energy saving: notEnergySaving state and energySaving state have been defined in 3GPP Rel-11 based on this, the complete energy saving mechanism includes two elementary procedures:

- Energy saving activation (change from notEnergySaving to energySaving state) and
- Energy saving deactivation (change from energySaving to notEnergySaving state)
The detailed procedures and conditions for energy saving activation and deactivation are described in [97].

Various Energy Savings Management (ESM) concepts can apply to different RATs, for example UMTS and LTE. However, 3GPP has specified that some of these ESM concepts may be limited to specific RATs and network elements, and specific solutions may be required for them.

Rel-11 describes three general architectures that are candidates to offer energy savings functionalities. The three general architectures are distributed, network management centralized and element management centralized. It also describes energy savings management use cases, the cell overlay use case, and the capacity limited network use case in detail. In addition to this, Rel-11 specifies the requirements for element management centralized energy savings and distributed energy savings and addresses the coordination issues between energy saving and cell outage. Signaling procedures and the exchange of measurement information for switching on/off different network elements is also discussed.

3GPP Rel-11 identified the need to coordinate various SON functions at different locations in the network. These functions can be related or dependent on each other or independent. The mode of operation between the SON Coordination Function and the SON Function, as well as the role of the SON Coordination Function, in the detection and attempt to resolve the conflicts, are specified in Rel-11. The possibility of conflict occurs when two or more SON functions try to change the same network configuration parameter for example, Ping-Pong modification of one configuration parameter between two or more SON functions. If two or more SON functions are dependent on each other the behavior of one may impact the behavior of other function. SON Coordination refers to preventing or resolving conflicts or negative influences between SON functions to make SON functions comply with an operator’s policy.

3GPP has defined how the Integration Reference Point (IRP) manager uses standardized capabilities to set the SON function(s) targets, and where their weights are needed for coordination of SON functions whose outputs are not standardized. For coordination of SON functions whose outputs are standardized, the context of optimization coordination is FFS. 3GPP has addressed the coordination between SON functions below Iff-N and Configuration Management (CM) operations over IfN-N. Examples of conflict situations are specified in Rel-11.

In a real network, conflicts may occur if centrally managed operations via If-N and several SON functions below Iff-N, running simultaneously, try to change same parameters at the same time in a short time period. Hence, coordination is necessary to avoid such conflict situations. To address this issue, 3GPP has identified a function referred to as a SON Coordination Function that will be responsible for preventing or resolving conflicts. The SON Coordination Function may be responsible for conflict prevention, conflict resolution, or both in parallel. It has been specified by 3GPP that such various SON Functions would be obliged to request permission to the SON Coordination Function before changing some specific configuration parameters. To summarize, the role of the SON Coordination Function in the detection and its attempts to resolve the conflicts along with the mode of operation between the SON Coordination Function and the SON Function have been specified in Rel-11. Some general principles and requirements, defining functions for minimizations of drive tests (MDT) have also been described in Rel-11. The MDT solutions should consider the constraints related to UE measurements and Location information. Also, Rel-11 has defined detailed mechanisms for
Management Based Activation, Trace Parameter Propagation, and Trace Record Collection in the case of signaling-based activation. Quality of Service (QoS) verification use cases beyond the coverage use cases addressed in Rel-10 has also been included in the Rel-11. To verify QoS, assess user experience from the RAN perspective and to assist network capacity extension, the MDT data reported from UEs and the RAN may be used.

3GPP Rel-12 mainly focus on inter-operability aspects for e.g. multi-vendor and inter-RAT SON functionalities of existing features as well as new features in some areas. The focus is also on evaluating different opportunities with more UE specific handling, and whether there are needs for new standardization support. For e.g. one possible concern is Ping-Pong handovers in case of different treatment of different UE types with different UE capabilities in two eNBs involved in load balancing. 3GPP Rel-12 will reach its functional freeze date in March 2015, which means no further functionalities and enhancements can be added in this release. The work is in progress for finalizing the 3GPP Release 13.

4.2.1 Multi-Vendor SON:
Multi-vendor SON operability have recently gained much attention, specifically in heterogeneous networks (HetNets) with macro and small cells deployment scenario. In a network, different network elements can be provided by different vendors to a given operator. In such a scenario, it is a challenge to make SON functions operate due compatibility issues for communicating between network elements and functions via multiple interfaces. To address this, 3GPP has defined a number of interfaces. In this section, only the interfaces relevant to the multi-vendor SON analysis will be discussed. Fig. 1 represents 3GPP Management Reference Model [98], [99]. It considers the cases of Centralized and Distributed SON where, Network Elements (NE) such as RNCs, eNB etc. are usually provided by the vendor together with the Element Manager (EM). The only interface that is standardized is the EM north bound (Type 2) interface. This interface is typically conformant with 32 and 28 series of 3GPP standard.

3GPP Management Architecture
Centralized SON and Distributed SON
X2 is a RAN control interface that is used to connect an eNB with its neighbor eNBs. Over this interface, usually two kinds of information are exchanged i.e. load or interference related information and handover related information. The X2 application protocol includes the support of some of the SON functions. Itf-N (a.k.a. Type 2) between the Element Manager (EM) and the Network Manager (NM) and Type 1 between the Network Elements (NEs) and the Element Manager (EM) are the management interfaces. The mapping of SON operations to certain management actions at the NE (RNC or eNodeB) is performed by the Element Manager (EM) over the Type 1 interface. The input data provides information such as performance measurements, fault alarms, notifications etc. The centralized NM-based SON manager monitors and analyzes this input data and optimize network with the help of algorithms. Once this is done, the appropriate actions to resolve problems at the affected network nodes are triggered.

The 3GPP management specifications focus more on Type 2 and slightly less on Type 1 management interfaces. This Type 1 interface can be proprietary as usually, the EMs are supplied by same vendor as eNBs. Type 1 interface will therefore not be used for multi-vendor interoperability.

4.2.2 SON Architecture Implementation in Multi-Vendor environment

There are a number of ways to implement different types of multi-vendor SON architectures considering the interoperability issues that may arise [100]. General network architecture to identify the possible location of SON algorithms and make general assumption on the potential interworking, based on the location of these functions is depicted in Fig. 2. In centralized SON architecture, SON may operate at the network management (NM) level via the Itf-B interface. Here, the element manager (EM) translates operations at Itf-N to corresponding actions at Type 1 interface. In case of distributed SON architecture, the SON functions may reside at the DM (Domain Manager) level or at eNB level, Itf-N or X2 can be used for the communications of the SON functions between vendors, depending on the SON function considered, and implementation choices. This includes the EM centralized architecture as described in [101].

There are 2 main architectures for implementing SON functions in a multi-vendor environment.

Architecture option-1, distributed SON for small and macro cells is presented in Fig. 2.
Figure 40: Distributed SON

Here, the architecture assumes that a given SON function is located at EM or eNB level for both vendors. In this case each vendor would have implemented standardized 3GPP X2 interfaces [102] and will be able to support most of the X2 exchanges for inter-working purposes.

The main challenges for vendor coordination over X2 in this architectural scenario are [103]:

- Alignment of support of 3GPP optional signaling to enable other vendor algorithms.
- Alignment of exchanged value meanings
- Alignment of timing for function monitoring and reporting.

The architecture option-2 is depicted in Fig. 3. In this case, it is assumed that a given SON function is located at EM or eNB level for Small Cell vendor (B) and at NM level for Macro Cell vendor (A).

In such an architecture, the small cell vendor is expected to have implemented standardized 3GPP X2 interface [102] and will be able to support most of the X2 exchanges for inter-working purposes, however Macro Cell vendor may or may not have supported SON related X2 signaling and/or IEs.
The main challenges for vendor coordination over X2 in this architectural scenario are:

- Alignment of exchanged value meanings
- Alignment of timing for function monitoring and reporting
- Alignment of support of 3GPP optional measurement reporting to enable other vendor algorithms.
- Support and alignment of X2 and Itf-N information exchange, which may not be standardized

Further details regarding interoperability issues and the recommended remedies are discussed in [99].


4.3.1 Introduction

One of the key challenges faced by the mobile network operators today is to support the growing demand for mobile data traffic in a cost effective manner [104]. In this direction, the installation of small-sized base stations into the macro cellular network layout, a.k.a. small cells, has recently drawn significant attention. The integrated cellular network infrastructure of macro cells and small cells is widely termed as heterogeneous network (HetNet). Small cells can boost the area spectral efficiency in the licensed spectrum and bring the cellular network closer to the end user in a cost-effective manner. The support of small cells is integral part of the Long Term Evolution – Advanced (LTE-A) system, which also enables the end users to deploy small cells in an unplanned fashion [105]. Among others, small cells feature edge-based intelligence that enables them to adapt their uplink (UL) and downlink (DL) transmissions in order to avoid cross-tier interference with the macro cell network and support the Quality of Service (QoS) requirements of the associated users. Under this viewpoint, Time Division
Duplex (TDD) is considered as a key enabler for achieving flexible and on-the-fly adaptation of the UL and DL resources at the small cells. In parallel, the mobile network operators have typically access to a fixed set of network resources, a.k.a. dedicated resources, which can be portions of the licensed spectrum or a set of physical network components, e.g. base stations. Aiming to avoid the underutilization of the physical resources and enable efficient resource sharing among multiple network operators, recently, there has been a surge of interest for leveraging the benefits of Software Defined Networking (SDN) in mobile cellular networks [106]. Among others, SDN can reduce network provisioning, enhance network flexibility, and open the road ahead for innovative, dynamic, and cost-effective solutions based on the concept of network re-programmability. Besides, such flexibility creates new business opportunities for the mobile operators and enables on-the-fly network-tuning with respect to the applications or services accessed by the users.

Recently, the industry has shown significant interest in the concept of co-existing FDD-TDD LTE-A network. SK Telecom, Korea Telecom and SingTel have demonstrated TDD-FDD carrier aggregation. In June 2014, SK Telecom used LTE-A carrier aggregation to combine 200 MHz bandwidth (9 TDD carriers and 1 FDD carrier) achieving a peak downlink throughput of 3.8 Gbps which is a world record. In October 2014, SingTel aggregated 20 MHz of LTE FDD spectrum with 20 MHz of LTE TDD spectrum to achieve a peak downlink data rate of 260 Mbps. Operators planning to deploy a converged LTE FDD 1800 MHz and TDD 2.3 GHz network are supported by 187 user devices (including operator and frequency variants) that support both band 3 and band 40 [107]. Operators Aircel and Bharti Airtel (India) are planning converged FDD-TDD networks, while Nokia Networks, STC showed first TDD-FDD aggregation with commercial chipset in Middle East & Africa (in Nov/December 2014) [108].

4.3.2 Related Work

In HetNets nowadays, the dense yet unplanned deployed of small cells results in unbalanced utilization of the physical resources among the cellular infrastructure. Even though a specific subset of cellular stations can be overloaded, other nearby cellular stations may underutilize their dedicated resources, due to the irregular spatiotemporal variations of the user traffic or the selected service provider. In such situations, Radio Access Network (RAN) and spectrum sharing can considerably improve the area spectral efficiency of the current cellular networks. For example, multiple mobile network operators can share their infrastructure so as to reduce the number of active network equipment, e.g. base stations, or reduce the capital investments for setting up the LTE-A RAN infrastructure. In two recent studies, 3GPP has overviewed the service and business requirements for realizing the so-called network sharing paradigm [109] and highlighted the required architectural enhancements [110]. Current literature includes a notable amount of frameworks, architectures, and mechanisms that allow for effective sharing of radio-resources or physical network components that can reside in either the access or the core network. The solution in [111] aims at providing on-demand infrastructure and spectrum sharing among different operators in 3GPP LTE networks. The authors propose mechanisms for virtualizing the evolved Node B (eNB) hardware by creating logically independent base stations, a.k.a. virtual eNBs (VeNBs). Among others, the proposed mechanisms are shown to balance the traffic load among the eNBs involved in the sharing process. The authors in [112] propose a multi-tenant solution that enables resource isolation and coexistence of independent policies among different eNBs’ instances. To achieve this, they propose a two-layer resource-scheduler composed of a global and a local resource scheduler that permits to implement different scheduling policies to different eNBs. A similar approach is discussed in [113], which modifies the scheduler of a shared eNB to isolate the traffic between multiple operators, while achieving a multiplexing gain. The key component in [113] is an entity called the
Hypervisor, which virtualizes the eNB into a number of slices allocating the Physical Resource Blocks (PRBs) among multiple virtual operators according to existing agreements. Although the results in [111]-[113] are promising, further work is required to adapt the presented solutions in a HetNet environment, where different functional capabilities are supported by the cellular base stations. An SDN-based framework is proposed in [114], to enable efficient on-demand sharing of base stations that belong to different operators. The key idea is to allow the cellular users to attach to the nearest base station. The OpenRAN architecture is introduced in [115], to leverage the convergence benefits of HetNets and achieve, at the same time, customization via network programmability. Network virtualization (NV) is realized at different levels including the application, spectrum, and network level, achieving a higher granularity of the combined pool of network resources. The use of SDN paradigm as a tool to simplify cellular network management is analyzed in [116], where a hierarchical architecture of a local and a global SDN controller is proposed. The local controller manages the processes inside a single network, while the global controller handles the events throughout the HetNet and coordinates the RANs through the backhaul. Different from [114]-[116], in this work we focus on the resource sharing problem in a multi-operator integrated FDD macro cell and TDD picocell LTE-A system, employing network programmability through re-configuration of the TDD frames at the pico eNBs. The problem of re-configuring the TDD frames of pico eNBs has also been considered in [117] [118] [119]. However, different from these approaches, in this chapter we additionally consider the amount of resources shared in the FDD system and the requirements of the users associated with the TDD picocells. Focusing on TDD systems, the idea of flexible spectrum sharing is investigated in [120]. The authors in [121] consider a mix of TDD and FDD, where the TDD occupies the guard band spectrum between the FDD UL and DL. Different from [120]-[121], in this work we consider elastic resource sharing in a multi-tenant multi-operator LTE-A HetNet environment.

4.3.3 Elastic Resource Management in LTE-A HetNets

In this section, we introduce the concept of elastic resource sharing in the FDD/TDD LTE-A HetNet (Section 4.3.1) and overview the main components of the proposed SDN-based solution for resource sharing among the TDD and FDD systems. Section 4.3.2 presents an innovative architecture for efficient resource coordination and sharing between the FDD system of macro eNBs and the TDD system of pico eNBs. In section 4.3.3, we present the proposed re-configuration algorithm for the TDD frames, which enables efficient resource sharing between TDD system of pico eNBs and corresponding FDD system of macro eNBs.

4.3.3.1 FDD/TDD Elastic Resource Sharing Concept

TDD and FDD operation is an integral part of the baseline functionality of the LTE-A system. Nevertheless, FDD is widely identified more suitable for applications that generate symmetric traffic, e.g. voice-centric services, while TDD for serving bursty / asymmetric data traffic, e.g. social media services and machine-to-machine (M2M) communications. In this work, we consider a heterogeneous network with overlaid FDD macro eNBs and underlying TDD pico eNBs that may belong to different network operators. All cellular stations are managed by a common infrastructure provider that leases the infrastructure to the FDD/TDD operators based on service level agreements (SLAs). The infrastructure provider also allocates the frequency bands to the FDD and TDD systems. The use of frequency bands for FDD and TDD transmission modes depend on the geographic region and the SLAs between the infrastructure provider and the network operators.
Here, we consider the scenario where the FDD (or TDD) system of a tagged macro (or pico) operator requires more radio-resources for efficiently supporting the ongoing services of the associated users. To address this requirement, we focus on the scenario where the TDD pico system is capable of leasing a part of its allocated resources to the FDD macro system in a highly efficient yet scalable manner. Accordingly, the FDD macro users can employ carrier aggregation to co-utilize the (potentially distant) set of radio resources of the TDD pico system. This transfer of radio resources, a.k.a. elastic resource sharing, can also take place towards the opposite direction, i.e. from the FDD to the TDD system. The FDD and TDD systems can co-exist and complement each other serving different types of traffic provided that interference mitigation is assured [117].

To achieve elastic resource sharing among the two systems, we consider the presence of a centralized SDN Controller, which acts as a resource brokering entity with global resource knowledge. In addition, aiming to efficiently handle the transfer of resources in the case of TDD system, we also consider that the TDD pico eNBs can re-configure the UL/DL frame ratio in the emitted TDD frames. The aforementioned process is performed in relation to the current resource share and the demands of the cellular users, while it allows for efficient re-adaptation of the TDD resources at the pico eNBs in a timely manner.

4.3.3.2 SDN-based Network Resource Management

Fig. 4 depicts the proposed SDN-based network management architecture. Without the loss of generality we focus on the scenario where radio-resources are leased from the TDD system to the FDD one, and consider a LTE-A HetNet of one macro eNB operator, coined as Operator A, and one pico eNB operator, coined as Operator Z. The macro eNBs of Operator A use FDD, the pico eNBs of Operator Z use TDD, while all types of eNBs are assumed capable of communicating with the central SDN controller (NV-aware eNBs). The discussion below can be readily extended to the scenario of multiple FDD macro eNB operators and multiple TDD pico eNB operators. The FDD and TDD network operators have to establish SLAs prior to the employment of elastic resource sharing. Apart from the dynamic sharing of radio-resources, the FDD and TDD network operators may also share the common base station infrastructure so as to enable their users to access the closest base station in proximity [114]. The latter functionality is termed as multi-tenant operation. Aiming to cover both these functionalities we extend the base station virtualization model in [114] by additionally enabling on-demand network reconfiguration capabilities for resource sharing of the underutilized spectrum resources between the FDD and TDD systems.

As shown in Fig. 4, the intelligence for resource management resides at the SDN Controller, which also provides application programming interfaces (APIs) for over-the-top (OTT) or business applications. Each transmission mode, i.e. FDD or TDD, is managed by a different control application that is capable of acquiring the knowledge of the network state by means of periodic information exchange with the FDD eNBs and TDD pico eNBs. The multi-tenant TDD pico-eNB architecture consists of a Hypervisor that is capable of virtualizing the physical resources to enable multiple operators to share the available bandwidth while remaining isolated from each other. To this end, two agents at the pico eNB are remotely assisted by two distinct manager applications at the SDN Controller: the Multi-Tenant Manager (MTM) and the Radio Resource Manager (RRM).
Figure 42: SDN-based network resource management architecture

More specifically the MTM is the place where the multi-tenancy policy resides and where the handover (HO) decision towards the TDD pico eNBs is taken. It is the entity that instructs an agent into the Hypervisor at the pico eNBs, referred to as Multi-Tenant Agent (MTA), providing it with real-time information for enabling efficient sharing of the multitenant pico eNBs into a prescribed number of slices. Note that a slice is represented by the list of users belonging to a tenant operator and the amount of resources that the multitenant pico eNB shares. Another agent, referred to as Resource Transfer Agent (RTA), is responsible for implementing the radio resource transfer procedure. To achieve this, a MAC scheduler agent cooperating with the RTA and MTA provides to the tenant operators an abstraction of the MAC layer. The MAC scheduler allocates to each operator a number of the available PRBs in a transparent way, i.e. transparent handling of all the operations related to the real-time configuration of each slice. The upper layers of the protocol stack are emulated in a pool of software applications, referred to as VeNBs, where each VeNB is managed by a different tenant operator and is logically connected with the core network of the tenant operator.

The logic signaling flow for both the multi-tenancy procedure and the network re-configuration for elastic resource sharing is illustrated in Fig. 5. The multi-tenancy policy, that enables HO of the users to multi-tenant pico eNBs, takes into account periodic physical measurements that give to the MTM, a global view of the network state. Once the HO decision is made, the MTM selects the more suitable target multi-tenant pico eNB and sends to the MTA a slice re-configuration request. It also initiates the HO procedure.
The communication between the SDN Controller and each NV aware base station is based on a combination of OpenFlow protocol [106] and an appropriate high layer protocol, i.e. based on UDP, as in [114]. The MTM uses OpenFlow to instruct the MTA agents sending them appropriate rules to dynamically configure each slice and enable the delivery of the packets from the users to the appropriate VeNB. Moreover, a high layer protocol is used to permit the exchange of messages between the entities involved in the network re-configuration procedure (RRM and RTA agents of the TDD pico eNBs). The RRM forecasts the resource availability and enables the transfer of resources between the TDD and FDD systems. To this end, the RRM is periodically informed about the network state, i.e. the bandwidth utilization of each system, by the RTA agents of both FDD and TDD NV-aware base stations. In such a way it can acquire a global knowledge of the network state and distribute the available resources in an efficient way.

The RTA agent of each NV-aware base station is responsible for collecting the bandwidth requests delivered by the tenant operators and distribute the available bandwidth among them. Such requests are sent to the FDD manager, if such request is performed by an FDD eNB, and forwarded them to the RRM module of the hosting system, e.g. the TDD manager. In this scenario, the TDD manager performs admission control and TDD frame re-configuration by the policy described in section 3.3.3 below. If the TDD manager is capable of supporting the requests for radio-resources from the FDD tenant operator, it transfers underutilized resources from the TDD pico eNB operator to the FDD macro eNB one. This operation is performed as follows: the TDD manager instructs the MTA agents of the involved TDD pico eNBs how to adjust the bandwidth availability so as to make part of its resources to the FDD macro eNB operator, while the FDD manager of the respective macro eNB operator allocates the acquired resources among the overloaded FDD eNBs.

4.3.3.3 Dynamic Resource Sharing Among FDD/TDD HetNets

When transferring or sharing resources among different operators there is a need for a mechanism to assure that the resource gain of a certain operator is not resulting on starving the users of the other. To regulate such resource transfer we adopted an approach, similar to the Distributed Fair Capacity Based Channel Allocation model elaborated in [122]. Without loss of generality, we focus on the scenario of transferring resources from the TDD operator to the FDD one. Such a process is based on the average
capacity gain considering the additional resources provided to the FDD macrocells against the capacity lost on the TDD pico eNBs from where the resources are taken. The actual capacity gain and the capacity loss are estimated using the modified Shannon formula included in [123]. In particular, the capacity gain is calculated considering the mean SINR of all users in the FDD macrocell, where additional resources are transferred, while the capacity loss is estimated on TDD pico eNBs, from where resources are borrowed, taking into account only the experience of the users with the worst SINR, i.e. the users having SINR < 0.

Let $C_{FDD,n}$ be the capacity gain for the FDD macrocell when it borrows a specified set of resources from the TDD pico eNBs. The average capacity increase $G_{FDD}$ for the FDD macrocell users is given by:

$$G_{FDD} = \frac{1}{N_{FDD}} \sum_{n=1}^{N_{FDD}} C_{FDD,n}$$

Where, $N_{FDD}$ is the total number of users in FDD macrocell. If the capacity loss in the TDD pico eNBs is less than the average increase in capacity gain for the FDD macrocell users, then the requested resources can be transferred, while assuring the desired user QoS for the TDD pico eNBs. In other words, the resources are transferred only if $G_{FDD} > L_{TDD}$, where $L_{TDD}$ is the capacity loss experienced by the worst SINR users in the TDD pico eNB region. It should be noted that the capacity gain and the capacity loss calculation for transferring resource in the opposite direction, i.e. from FDD to TDD, follows a similar process.

We envision that the resource transferring condition described above assist the SDN Controller to take the decision regarding resource sharing among different operators provided that the amount of desired resources, i.e. PRBs, is communicated or estimated by the SDN Controller. Communicating the desired resources could easily be accomplished via the north bound API from where application providers can ask the SDN Controller regarding particular QoS, e.g. capacity. Alternatively, the FDD macrocell upon starting experiencing congestion can signal the SDN controller requesting additional resources, which can be provided by exploiting the maximum possible amount of resources that can borrowed from the overlapping TDD pico eNBs. Since the SDN Controller has knowledge of the network state, i.e. load associated with macrocells and pico eNBs and the respective interference levels, it can estimate the maximum amount of resources it can transfer from the TDD pico eNBs that can satisfy the capacity gain and capacity loss resource transfer condition.

The SDN controller may also have knowledge of the specific TDD UL and DL resource utilization and hence can further optimize the UL/DL ratio associated with particular TDD pico eNBs with the objective to increase the amount of resources that can be transferred towards the FDD macrocell. The rational here is to determine an appropriate UL/DL frame with respect to the residing user UL/DL demands, which can free more resources to be transferred towards the FDD macrocell. The mechanism for determining the corresponding UL/DL frame can be based on the same heuristics as the ones used in [120]. It should be noted that the amount of resources that can be shared among the FDD and TDD operators may vary depending on particular QoS demands and traffic conditions.
4.3.3.4 System-Level Simulation Results

This section presents some preliminary system level simulation results for assessing the performance gains attained from the proposed SDN-based framework for elastic resource sharing in LTE-A HetNets. We consider an LTE-A HetNet composed by of FDD macro eNBs and TDD pico eNBs of a different operator. Two different operators are assumed to operate the macro and the pico eNB infrastructure, respectively. We focus our analysis to the scenario where the TDD pico eNB operator is willing to share part of its resources to the FDD macro eNB operator, depending on the resource availability in both systems. The elastic resource sharing is performed based on SLAs that defines the maximum bandwidth that can be transferred from the TDD pico eNB operator to the FDD macro eNB operator and vice-versa.

In more detail, we consider a LTE-A HetNet with seven FDD macro cells and seven pico eNBs that are uniformly dropped within the FDD macro area with a minimum separation distance of 120m. The users of both the FDD macro and TDD pico eNB systems are uniformly distributed within the cellular coverage, whereas two different types of traffic are assumed between them. The FDD macro users are considered to have a H.264 video flow encoded video streaming traffic at 440 kbps, while the TDD pico users host a Poisson-distributed FTP traffic that is modeled in line with the methodology in [124]. The macro eNBs are assumed to utilize the EXP scheduler, which gives priority to real time DL packets which are buffered for more than a target delay threshold of 0.1 sec [125]. On the other hand, the traffic at the pico eNBs is randomly and independently generated in the UL and DL directions. Since we are not interested in the impact of user mobility in the TDD pico eNB system, we consider uncorrelated slow fading. The path-loss model for both systems are adapted based the 3GPP case 1 model defined in [124]. The remainder parameters of our simulation model are summarized in Table I.

Table 10: System level simulation parameters

<table>
<thead>
<tr>
<th>Basic Radio Configuration Parameters [19,20]</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Macro eNB Inter-site distance</strong></td>
<td>500m</td>
</tr>
<tr>
<td><strong>Pico eNB minimum separation distance</strong></td>
<td>120m</td>
</tr>
<tr>
<td><strong>Shadowing standard deviation</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Spectrum Allocation</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Max Tx Power [dBm]</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Antenna Gain [dB]</strong></td>
<td></td>
</tr>
<tr>
<td><strong>UE-Pico eNB Path Loss</strong></td>
<td></td>
</tr>
<tr>
<td><strong>UE-Macro eNB Path Loss</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Macro eNB Fading</strong></td>
<td></td>
</tr>
</tbody>
</table>

- **Macro eNB Inter-site distance**: 500m
- **Pico eNB minimum separation distance**: 120m
- **Shadowing standard deviation**:
  - Macro Cell: 8 dB
  - Pico Cell: 10 dB
- **Spectrum Allocation**:
  - Macro Cell: 10 MHz UL/DL
  - Pico Cell: 10 MHz UL/DL
- **Max Tx Power [dBm]**:
  - Macro: 46
  - Pico: 30
- **Antenna Gain [dB]**:
  - Macro / Pico: 15 dB / 5 dB
- **UE-Pico eNB Path Loss**: $140.7 + 36.7 \log_{10}(R)$, $R$ in km
- **UE-Macro eNB Path Loss**: $128.1 + 37.6 \log_{10}(R)$, $R$ in km
- **Macro eNB Fading**: Jakes model
- **Macro eNB**: EXP Rule-target delay=0.1 s
Since the TDD re-configuration scheme is used for optimizing the UL/DL ratio in the TDD pico eNB system only, in Fig. 6 we plot the delay cumulative distribution function (CDF) of the H.264 video traffic in the DL of the FDD macro eNB system under two different schemes. The first scheme, coined as the FDD baseline scheme, corresponds to the performance of the macro eNB system without using the proposed SND-based framework. On the other hand, the FDD with SDN scheme corresponds to the performance of the FDD macro eNB system under the proposed SDN-based framework. As expected, above the target delay threshold of 0.1 seconds, the performance of both schemes the baseline and the SDN-based approach is the same. However, a notable delay gain is observed for the SDN-based scheme below this threshold, i.e. the CDF delay of the SDN-based scheme is higher to that of the baseline scheme. Interestingly, the employment of SDN-based elastic resource sharing reduces the H.264 traffic delay in the DL direction by up to 21% compared to the baseline scenario where no SDN based sharing is applied. This improvement follows from the utilization of additional resources provided by the TDD pico eNB system.

Figure 44: Application Layer Delay CDF of the FDD Macro eNB System

Although the reduction of the application layer delay in the FDD macro eNB system is prominent (Fig.6), the negative impact of utilizing less radio-resource in the TDD pico eNB system, i.e. the ones that have been shared with the FDD macro eNB system, should also be investigated.
In Fig. 7, we plot the application layer delay CDF of the TDD pico eNB system under three different schemes: a) no sharing of resources with the FDD system and no re-configuration of the TDD frames (blue line), b) SDN-based elastic resource sharing with the FDD system and no re-configuration of the TDD frames (green line), and c) SDN-based elastic resource sharing with the FDD system and re-configuration of the TDD frames (red line). In the first two schemes the reference UL/DL configuration-1 (60% downlink and 40% uplink) has been considered for the TDD frames at the pico eNBs [117]. In the third scheme, we consider an UL/DL reconfiguration timescale of 10ms and the seven UL/DL configurations available for TDD-LTE [117].

As shown in Fig. 7, the sharing of resources with the FDD macro eNB network degrades the performance of the TDD pico eNB system, if a static UL/DL configuration is applied, i.e. compare the blue and the green lines. In the contrary, if the elastic resource sharing between the two systems is combined with the dynamic re-configuration of the TDD frames on a per pico eNBs basis (red line), the employment of the proposed SDN-based framework is shown to attain notable performance gains for the TDD pico eNB system as well. This performance improvement mainly follows from the efficient adaptation of the UL/DL ratio with respect to the ongoing user services and the resource availability in the TDD pico eNB network.
5. Towards the standardization and resource sharing optimization among D2D communications

5.1 Resource sharing over D2D: introduction and state of the art

The International Telecommunication Union – Radio communication sector (ITU-R) has initiated activities on new issues arising in the context of the International Mobile Telecommunications-Advanced (IMT-Advanced) recommendations. One of the prominent topics considered is the concept of Device-to-Device (D2D) communications, i.e., direct communications in a cellular network, without the intervention of the Base Station (BS), when the transmitter and the receiver are in close proximity. Differing from conventional approaches, such as Bluetooth and WiFi-direct, D2D communications utilize licensed spectrum, while no manual network detection-selection is needed. Comparing to the very appealing cognitive radio communications, where secondary transmissions are allowed in parallel with primary cellular transmissions, D2D communications are established by standard/primary cellular users, reaping the benefits of being synchronized and controlled by the central (primary) BS.

D2D communications may be classified using various criteria, as illustrated in Figure 46. Depending on the type of spectrum used, we distinguish inband and outband D2D, which utilize licensed and unlicensed spectrum, respectively [126]. Furthermore, D2D may work as an underlay to the standard cellular operation by reusing resources with standard cellular users, unlike the overlay mode, where specific dedicated resources are either statically or dynamically assigned exclusively for D2D operation. Regarding the level of control of the operator in the D2D setup procedure, we find autonomous and controlled schemes. Moreover, regarding the initiation of the D2D communication request, there are two options: Either the D2D request is fully transparent to the end-user, whose communication is automatically switched from cellular to D2D mode by the operator (network-originated), or the D2D mode is originally (explicitly) requested by the user (user-originated). Finally, D2D transmissions may be unicast or multicast/broadcast, where the former case describes peer-to-peer links for direct communication or relaying links (e.g., for coverage extension purposes), while the latter would be more appealing for social and commercial applications, such as proximity-based advertisement or public safety scenarios.

![D2D classification types](image)

The introduction of D2D communications in cellular networks is beneficial from a variety of perspectives. The short distance between D2D users results in better channel conditions, leading to higher data rates, lower delays and lower energy consumption. Additionally, D2D users are connected through a direct link and the intermediate transmission to a BS is avoided, saving network resources and
processing effort from the network. Also, the coexistence of cellular and D2D links can lead to more efficient spectrum utilization and higher spatial spectrum reuse, while new business models, probably with a new charging policy for users, may be designed.

However, D2D communications do not come without a cost. On the one hand, interference-free conditions between D2D and cellular transmissions, as well as among D2D pairs are required. In this direction, a variety of resource management approaches can be found in the literature [127]-[131]. A key challenge is the design of mechanisms that inform the BS about the channel conditions among all nodes, either directly in the form of periodic measurements of the ongoing communication quality [127]-[128], or indirectly via neighborhood detection [129]-[130]. An efficient design option is the exploitation of the uplink cellular period, where the cellular interference victim is the immobile BS [131]. On the other hand, the efficient management of the D2D links is essential, including the peer discovery (D2D receiver discovery), the D2D connection establishment and maintenance, and the specific changes that must be made to the coexisting cellular network.

Recognizing the importance of these tasks, the interesting idea of switching a cellular connection to a direct (D2D) one and vice versa, taking into account performance criteria, is discussed in [127],[131]. However, this technique consumes operational resources of the core network, while the D2D connections are mainly used for network performance optimization and not for spatial spectrum reuse optimization. Also, the cell is inevitably loaded with extra control signaling. In addition to that, when thinking of D2D communication, we should not neglect the human factor; for instance, two users are aware that they are located in the same building during working hours; hence they can definitely reach each other via D2D directly. In this case, an indirect connection (e.g., cellular) would be costly and redundant.

From a more theoretical perspective, analytical results guarantee that D2D communications can be an important tool for enhanced network capacity and spatial spectrum reuse [132]-[133]. Unfortunately, these results are not reflected in the present standardization field. D2D communications for peer-to-peer communications are absent from the specifications of recently standardized cellular systems, such as the Long Term Evolution (LTE) [134]. In LTE Release 12, D2D communications are mainly examined under the perspective of emergency and public safety. In section 5.3 below de discuss the issue of enabling D2D communications in LTE, motivated by the previous discussion, and taking into account the current standardization. This discussion is based on [135].

The feasibility of having D2D as an underlay in LTE-A networks is discussed in [136]. Identified are the necessary enhancements to an LTE-A network in order to integrate the functionality of D2D communications. The D2D session setup and management procedure are described in the frames of the System Architecture Evolution (SAE) architecture. One option for detecting and enabling D2D communications is by earmarking the local traffic at the Packet Data Network (PDN) gateway and triggering the eNB to check if the D2D mode is more beneficial. The appropriate measurements are taken by the UEs themselves and sent to the eNB. A different approach is also proposed, where UEs with specific SIP addresses explicitly require the D2D operation mode. In this case, a new light SIP handler needs to be added to the Mobility Management Entity (MME) functionality. Furthermore, the interference coordination problem both in uplink (UL) and downlink (DL) are studied. The control and limitation of the UE's maximum transmit power provides a means to handle this issue. Finally, the feasibility of D2D communications as an underlay to LTE-A is analyzed. Simulation studies in an indoor, fully loaded and interference limited environment show that the co-existence of D2D and cellular communications increases the overall throughput of the system.
Similarly, the authors in [137] discuss all the necessary enhancements in terms of new mechanisms or functional blocks that need to be added to an LTE-A system in order to enable D2D communications. D2D communications are appealing for data intensive short range communications, thus new service opportunities arise. Some of the proposals in order to support this new type of communications are: a D2D radio bearer needs to be setup with the assistance of the MME, UDP/IP can be used for information exchange, interference coordination and link adaptation mechanisms need to be adopted and channel measurements should be made and sent to the eNB (e.g. Sounding Reference Signal - SRS). Moreover, the authors explain that D2D communications are expected to have a short range and to offer limited mobility support. Simulations in both an indoor and outdoor (Manhattan grid) environment show that there is an important trade-off between the SINR degradation of cellular links and the achieved D2D range. Nevertheless, D2D communications with practical range are feasible, especially for the less challenging scenario of reusing UL resources only.

Regarding the problem of resource allocation in D2D communications, a variety of schemes can be found in the literature (e.g., [128]-[129], [138]-[140]). The majority of them adopts the underlay mode and proposes schemes that guarantee low-interference levels at cellular and D2D receivers. However, the interference-free coexistence of D2D and cellular communications is a quite challenging task, leading to limited spatial spectrum reuse. On the other hand, the overlay mode promises a high number of concurrent D2D links without affecting the cellular communications. In both cases, current solutions propose resource allocation schemes where interference or topology information is available at the base station [128],[139]. Under this perspective, an efficient approach is to represent the interference/topology information into a graph, as described in [129],[140]. In section 5.2 below, we describe one such approach in that direction based on [141]. Moreover, next we describe more resource allocation approaches found in the literature.

For instance, the authors in [142] try to optimize the resource allocation to an underlay D2D network by formulating a Mixed Integer Nonlinear Programming (MINLP) problem. They propose a greedy heuristic Resource Blocks (RBs) selection algorithm and simulate the system under very realistic assumptions concerning the LTE network architecture, the radio access technology, the channel pathloss and shadowing, the scheduling algorithms and the interference conditions. The decisions on resource allocation are based on the observation that the performance is improved when the channel gain between the cellular UE and the D2D device or between the eNB and the D2D device for downlink and uplink, respectively, is less. The results show that the harmful interference is reduced compared to a random D2D RB assignment and the sum network throughput increases while more and more D2D links are setup. The work also assumes that an intercell interference mitigation scheme works on top of the proposed scheme and that the eNodeB is fully informed for the SINR among the D2D and cellular links.

In [143], the authors present a spectrum sharing model, where D2D users reuse the resources from Cellular Users (CUs) and they manage to derive identical analytical and simulation results regarding the probability of link existence in case of single or multi-channel environments. In the proposed model, the D2D users are able to determine their necessary transmission power themselves, by estimating the channel pathloss via the received signal from the BS (whose transmission power is considered known), with respect to the minimum required BS SINR. The end-to-end paths are discovered using the ad-hoc Dynamic Source Routing (DSR) discovery protocol. The authors have shown that the probability of link increases when the number of channel increases, when clusters are introduced in the cell area, when the
pathloss exponent is larger (i.e. interference decays faster) or in the case of multihop D2D communications when the number of idle users is larger.

Furthermore, in [144], a hybrid network of cellular and underlaying D2D users is studied, where resources are being reused in the DL by using a labeled time slots scheme. The proposed mechanism consists of a handshaking procedure and a resource allocation procedure. During the former, using CSMA/CA-like logic (RTS/CTS messages) the D2D transmitter and D2D receiver create a direct logical link to each other. The resource allocation procedure is based on the dissemination of signaling information within a Common Control Channel (CCCH), the reporting to eNB of interference information caused to cellular users by D2D users, and the allocation of shared or dedicated timeslots to cellular and D2D users, according to the near-far interference risk due to parallel D2D communications. Simulation results in terms of system throughput show a relatively satisfying performance.

In [127] the authors demonstrate the coexistence possibility of D2D and cellular communications inside an interference-limited network, when both connection types share the same UL resources. A BS is responsible for initializing the D2D links and allocating the OFDMA RBs, thus possessing the knowledge to control the intracell interference. A UL D2D power control scheme is introduced, which uses parameters such as the "backoff value" of the D2D transmit power and the "power boosting factor" of the UL transmitter. Considering that priority is given to the cellular communications, this UL-based power control scheme results in tolerable interference to the primary network as well as to satisfactory D2D link performance. Moreover, various resource allocation modes are studied and compared: DL-, UL-, separate resource sharing and cellular mode sharing (traditional system) based on the sum rate of cellular and D2D connections. Simulation shows how beneficial D2D mode is for the system, measured as the rate ratio of the best sharing option and the “CellMod” scheme for various distances between D2D users and the BS. Finally, it has been demonstrated that indoor D2D communications do not harm outdoor cellular ones when sharing downlink resources.

The authors in [145] firstly present a short survey of D2D communications, as another add-on feature to LTE-A networks, which targets to optimize the spectrum utilization. The main challenges and research issues related to D2D are discussed (such as mode selection, transmission scheduling, power control, resource allocation, interference cancellation etc.) and the state of the art on this area is briefly given. In the second part of this work, a novel resource allocation scheme for D2D communications is proposed, that jointly performs Resource Blocks (RB) scheduling and Power Control (PC). The goal of this scheme is to increase the spectrum reuse by minimizing the transmission length spent for D2D communications (number of timeslots spent to satisfy the current D2D traffic demand). The resource allocation problem is solved using the “Column Generation method”, which tries to estimate the most efficient feasible access pattern (=subset of simultaneous D2D links on a RB), while guaranteeing adequate QoS to D2D users and restricted interference to cellular users. Simulations show that, indeed, this method decreases the transmission length of D2D links, with an inevitable parallel increase in the total power consumption.

The authors in [146] propose a mode selection scheme for resource allocation to underlaying D2D communications among three possible options: “Reuse mode”, where all resources are shared by both cellular and D2D links, “Dedicated mode”, where D2D communications are direct and they are fairly assigned unique resources and finally conventional “Cellular mode”, where D2D communications are relayed by the eNB. This proposed scheme selects for UL and DL that mode that provides the highest total system throughput, guaranteeing in parallel a minimum rate for the prioritized cellular links. The mode selection strategy takes into account the cellular and D2D link quality, the interference conditions
and the load situation of the cell. The authors have noticed a trade-off between the cell throughput and the D2D throughput ratio, for instance when using shared resources, even though the achieved cell throughput is very high, the probability of outage increases as well. Simulations in a multi-cell environment show that with the mode selection scheme the D2D throughput as well as the total cell throughput can be both kept at high levels. It is observed that in UL the reuse mode is dominant (as eNB is the only victim receiver), whereas in the DL the cellular mode is preferred, especially close to the eNB. Nevertheless, the signaling load for supporting such a mode selection scheme is significant.

Furthermore, the authors in [130] propose an interference cancellation mechanism in cellular hybrid networks, when reusing uplink resources. They model an LTE FDD system where all UEs listen to a common control channel that broadcasts relevant information (C-RNTIs, Cell-ID, power information etc.). The proposed resource allocation scheme is based on the idea that a cellular UE can detect a D2D neighbor. Then it reports this event to the eNB that collects this information from all cellular users and broadcasts it to the common channel. D2D users then self-allocate the resources, respecting this received interference-related information. The authors have shown that using this mechanism, the average D2D subsystem as well as the overall hybrid system throughput are significantly increased. Finally, they relate the system throughput to the SINR threshold considered when cellular users detect their D2D neighbors (the higher the threshold the lower the throughput).

The authors in [147] address the problem of power control optimization for three different resource allocation modes: when cellular and D2D users share non-orthogonal resources, when each type has orthogonal dedicated resources and when the BS acts as a traditional relay node. The resource allocation mode is selected on the criterion of maximizing the sum rate, separately for the UL and the DL. The power optimization problem is studied with greedy sum-rate maximization where cellular and D2D communications have the same priority and with rate constraints where cellular communications have priority. In the latter, a minimum rate is guaranteed for the cellular user. Simulations in a single cell, under the assumption of full Channel State Information (CSI) feedback to the BS, show that the averaged rate ratio between sum rate obtained from the best resource allocation and the rate obtained from the single cellular user case is higher for the greedy optimization. However, this mechanism significantly increases the cellular service outage probability, which is zero for power optimization with prioritization and rate constraints. Finally, the averaged cellular user rate is higher for this second optimization scheme.

Finally, D2D has been studied as an offloading mechanism. Specifically, the authors in [148] propose D2D as an appealing alternative to Local IP Access (LIPA) and Selected IP Traffic Offload (SIPTO) for cellular data offloading. They thoroughly describe a D2D communication path establishment procedure together with the necessary entities and enhancements in an LTE-A network. The major enhancements that need to be introduced for enabling operator-controlled peer discovery and D2D communication are: 1) the introduction of a new interface between D2D-enabled UEs, 2) the Proximity Service Control Function (PSCF) entity in the PDN-GW for detecting potential D2D traffic 3) a new D2D bearer (consisting of a packet filter and a radio bearer) for direct data exchange offloading the standard Evolved Packet System (EPS) bearer and 4) the ProSe Management (PSM) entity in the MME and UEs for activating/deactivating the D2D bearer context. Switchback mode to a cellular mode of communication is possible, and for this reason EPS bearers may be reserved (negative).
5.2 Ant Colony Optimization for Resource Sharing among D2D Communications

In this section, we focus on inband, overlay and unicast D2D links that are user-originated and controlled by the operator. We study the problem of D2D resource allocation in LTE-A networks, targeting at minimizing the amount of spectrum required for serving a specific number of overlaying D2D requests. The proposed scheme exploits the Ant Colony Optimization (ACO) theory and a graph representation of D2D mutual interferences, as a means to guarantee multiple concurrent D2D transmissions with specific target outage probability.

5.2.1 System Model

5.2.1.1 Problem statement

We consider an LTE-A network with availability for D2D communication. We refer to the LTE-A base station as evolved NodeB (eNB) and to the cellular end-user as User Equipment (UE). The eNB is responsible for allocating spectrum resources to the UEs every scheduling interval of 1 millisecond, known as the Transmission Time Interval (TTI). Regarding the resource allocation procedure, the available spectrum is divided into allocation units referred to as Resource Blocks (RBs).

After the scheduling process, a number of D2D requests are assigned for transmitting in a specific TTI. Sequentially, per TTI, a resource allocator decides on how many and which RBs will be used by each one of the scheduled D2D requests. The major challenge for the resource allocator is to exploit current spatial spectrum reuse opportunities towards minimizing the number of RBs required for satisfying all D2D requests.

The spatial reuse of the same RBs by multiple D2D pairs may be possible due to the low range of the D2D transmissions. To elaborate on this, we illustrate in Figure 47 the case where a specific spectrum portion is utilized by uniformly distributed D2D transmitters in a cell (Figure 47a). All D2D users use fixed transmit powers that enable short-distance D2D links. Under this deployment, the summated signal strength in each grid point of the cell area is depicted in Figure 47b, identifying the feasibility of spatially reusing the spectrum multiple times in the cell area (locally-restricted interference is observed).

Figure 47: a) A snapshot of D2D transmitters' locations, b) The corresponding total interference inside a cell with multiple concurrent D2D transmissions.

Let $D$ be the number of D2D requests scheduled in a specific TTI. To minimize the spectrum resources required for serving those requests, the eNB should estimate the maximum number of parallel D2D transmissions that can concurrently use a specific spectrum portion. In other words, the problem under consideration is the following: *For a specific set of D2D requests ($D$) find the largest subset ($L$) that may
spatially reuse the same spectrum portion under constrained interference conditions, which ensure acceptable quality for the D2D links.

For instance, given the topology of Figure 48, where totally 9 D2D requests are present \( (D_1 = 9) \), the largest subset found is \( L_1 = 4 \). Hence, 4 transmissions (solid links) may safely reuse the same spectrum portion. The remaining 5 requests define a new set, \( D_2 \), for which the problem should be solved again, by finding a new subset \( L_2 \subseteq D_2 \). Here, we assume \( L_2 = 3 \) (dotted links). By continuing this iterative procedure, we eventually manage to estimate the minimum amount of subsets, denoted by \( S \). Otherwise put, we manage to find the largest possible subsets \( L_i \subseteq D_i \), \( i = 1, 2, ..., S \), with \( \bigcup_{i=1}^{S} L_i = D \).

![Figure 48: System model example, \( D = 9, S = 3, L_1 = 4, L_2 = 3, L_3 = 2 \).](image)

Since the exhaustive search for finding the optimal solution regarding these subsets is computationally very complex and practically inapplicable, we exploit the ACO theory, seeking for an adequate approximation. The first step for applying ACO is to represent the interferences between the D2D pairs into a graph, as explained next.

### 5.2.1.2 Graph representation of D2D interferences

The eNB initially creates an internal representation of the mutual interference conditions among concurrent D2D links (as if all are allocated with the same resources), in the form of a fully connected weighted graph. This graph is created by exploiting information provided by the UEs to the eNB. Specifically, all UEs report information about the received powers experienced due to any potential parallel D2D transmission (i.e., interference). Regarding the acquisition of this interference information by the UEs, this may be possible by using techniques such as the D2D peer discovery procedure, not investigated here, though. The eNB transforms the acquired interference values into a graph, as explained next.

Let \( G = (V, E) \) be a graph where the vertexes represent the D2D requests: \( V = \{v_1, v_2, ..., v_D\} \), and the edges represent mutual interferences between any two competing requests: \( E = \{(v_1, v_2), (v_1, v_3), ..., (v_1, v_D), (v_2, v_3), ..., (v_{D-1}, v_D)\} \). The edges are weighted and these weights represent the level of mutual interference between any two potential parallel D2D transmissions, dependent on the actual channel conditions. In order to quantify this, we introduce the *Interference Level Indicator (ILI)* term. ILI takes values according to a configurable interval scale, ranging from a minimum to a maximum value. The min/max values in this scale correspond to the min/max estimated interference in the current topology, while all in-between values are calculated uniformly. Hence, a [0-1] scale implies that interference is present or not, while a [1-100] scale is able to describe 100 distinct levels of interference. The mapping of interference values to ILIs is done by the eNB. Ignoring, the zero values, the produced graph is fully connected, because all vertexes are potentially able to interfere to
each other, either heavily (edge weight = large $ILI$) or almost negligibly (edge weight = small $ILI$). Therefore, in this representation, the lower the weights among nodes, the less the mutual interference, and consequently, the higher the chances for a safe spectrum sharing.

An example of a network graph representation is given in Figure 49, where nodes A-E represent five D2D requests and $ILI$s range from 1 to 15. Here, the $ILI$ between vertexes A and B is 14, implying that these parallel D2D requests are in close proximity, while between A and E it is 1, implying that these are very far from each other and probably not interfering at all.

![Figure 49: Interference graph of the D2D pairs.](image)

5.2.1.3 Probability of D2D communication outage

The communication of one D2D pair is considered successful, if the receiver’s Signal to Interference plus Noise Ratio (SINR) is above a pre-defined threshold, which guarantees acceptable quality; otherwise an outage occurs. As interference, we consider the summation of all received powers due to the rest of the D2D requests scheduled simultaneously, which belong to the same subset $L$. This SINR value is calculated inside the eNB for each D2D link by exploiting already available interference information. This is feasible since the eNB is already aware of the received power at one D2D receiver by any other parallel D2D transmission (information already used for the graph representation), be it either interference or the desired signal strength.

The finding of the maximum possible size of $L$ is driven by an iterative procedure inside the eNB. Measuring the number of outages out of all the scheduled D2D requests that compose $L$, the eNB gets an estimation of the average Outage Probability ($P_{out}$), which, in turn, is compared to a maximum acceptable threshold. If $P_{out}$ is less than this threshold, then the selected subset will have the current size of $L$; otherwise, the considered size of $L$ has to be decreased by one and the new $P_{out}$ has to be estimated. As for the finding of the optimal requests composing $L$, this is driven by the ACO procedure, described next, as an effort to avoid the exhaustive examination of all possible combinations of sets of size $L$.

5.2.2 Ant Colony Optimization

5.2.2.1 ACO for D2D resource sharing

When $L_{len}$, i.e., the subset length is given, the optimization problem is expressed as the minimization of a sum of weights:

$$\min \sum_{i=1}^{L_{len}} W(v_i, v_j), \text{ where } v_i, v_j \in V$$

where $W$ is a $DxD$ matrix of all $ILI$ weights. Thus, the problem is translated to finding which vertexes should be selected as a part of the solution $L$, minimizing a total cost. Here, the cost is the summated
interference experienced by the D2D pairs when they share the same RBs, expressed via the summation of the respective $I_{LI}$ weights. It is worth noting, that the diagonal elements of the matrix $W$ represent the channel conditions among a D2D transmitter and its target receiver (desired signal), hence they represent the quality of the D2D link.

Eventually, the problem of finding the vertexes composing $L$ becomes a redefined graph coloring problem, where two adjacent nodes can be colored the same, if they consist part of the same solution, whereas in the traditional graph coloring problem, two adjacent nodes cannot be colored the same. In our scenario, when two D2D requests (vertexes) consist part of the same solution, it means that they have been allocated the same RBs (colored the same) for D2D transmission purposes.

This problem may be solved using the ACO theory. According to [149], “The main underlying idea, loosely inspired by the behavior of real ants, is that of a parallel search over several constructive computational threads based on local problem data and on a dynamic memory structure containing information on the quality of previously obtained result”. ACO was originally designed for solving routing problems, namely, as a way to find optimum routing paths. Nevertheless, here we adopt this theory for solving a resource allocation problem using a network graph representation. The eNB is the network entity in charge of running the ACO algorithm.

Ants are treated here as “colored agents”, and each one is carrying a unique color. When an ant visits a node it “paints” it with this color. In the resource allocation scenario, this means that the ant allocates the same spectrum portion (color) to all visited nodes (i.e., D2D requests), which will then consist a part of this ant’s solution $L$. The main idea behind this ACO version is to exploit $I_{LI}$ awareness to estrange two D2D pairs with good mutual channel conditions (i.e., in close proximity). Overall, the redefined parameters of the ACO are:

- $i, j =$ D2D requests, or simply, the graph’s vertexes $V$.
- $n =$ the index of an adjacent node $j$, $\{j_1, j_2, ..., j_n, ...\}$.
- $N =$ the total number of ants.
- $k =$ the index to an ant.
- $L^k = \{v_i, v_j, ..., v_{k_{len}}\} =$ the solution of each ant $k$.
- $L_{len} =$ the length of this solution (subset length).
- $tabu^k =$ the “tabu list” of ant $k$, to prevent it from coloring the same node more than once.
- $r =$ the evaporation rate. It represents how fast the topology changes and hence, how fast acquired knowledge by past ACO iterations fades.
- $d_{ij} =$ the “distance” between two competing D2D requests, here equivalent to the $I_{LI_{ij}}$ between a D2D transmitter and a victim D2D receiver (higher $I_{LI}$ means higher distance).
- $\eta_{ij} =$ the attractiveness of moving from node $i$ to $j$. It indicates the a priori desirability of the ant’s next move, i.e., here, the desirability of assigning the same color to request $j$, provided that it has been already allocated to $i$.

$$\eta_{ij} = \frac{1}{d_{ij}} = \frac{1}{I_{LI_{ij}}}$$
• $C^k$ = the estimated cost that derives from ant $k$’s route. Here, it is the total ILI “experienced” by this ant, i.e., the sum of all ILIs of all edges belonging to the solution of this ant. In Figure 49, this cost would be 8+10+6, if the ant visited the nodes A-D-C-E.

• $\Delta \tau_{ij}^k$ = the pheromone deposited by ant $k$’s $i \rightarrow j$ move:

$$\Delta \tau_{ij}^k = \begin{cases} \frac{1}{C^k}, & \text{if both } i \text{ and } j \text{ were visited by ant } k \\ 0, & \text{if not} \end{cases}$$

• $\tau_{ij}$ = the trail level or amount of pheromone deposited for moving from $i$ to $j$. It indicates how proficient has been in the past the coloring of $j$ given the same coloring of $i$ and, thus, indicates the a posteriori desirability of this move.

$$\tau_{ij}(t) = (1 - r)\tau_{ij}(t-1) + \sum_{k=1}^{N} \Delta \tau_{ij}^k$$

• $\alpha$ = the level of importance of $\tau$, $1 \geq \alpha \geq 0$.

• $\beta$ = the level of importance of $\eta$, $\beta \geq 1$.

• $p_{ij}^k$ = the transition probability that ant $k$ will move from state $i$ to state $j$. It depends on both $\eta$ and $\tau$.

$$p_{ij}^k = \frac{\tau_{ij}^\alpha \eta_{ij}^\beta}{\sum_{\text{all feasible } j \text{ states}} \tau_{ij}^\alpha \eta_{ij}^\beta}$$

5.2.2.2 Algorithm flow

Having provided the terminology and basic idea behind the application of the ACO theory for this resource allocation problem, below we describe the proposed algorithm using pseudo-code.

Algorithm: ACO-based D2D Resource Allocation
5.2.3 Evaluation Results

For the purposes of evaluation, the LTE link level simulator of [150] has been exploited in order to setup the system environment with D2D communication available, as well as to reliably estimate the received signal strengths from potential parallel D2D transmissions, required as input by the ACO algorithm. For the simulation, the selected input parameters were compliant with LTE standardized values and ACO recommendations [151], as depicted in Table 11.

Table 11: Basic simulation parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System model</strong></td>
<td></td>
</tr>
<tr>
<td>Topology</td>
<td>1 square cell of 100m edge</td>
</tr>
<tr>
<td>Number of D2D requests, ( D )</td>
<td>Configurable ([5..25])</td>
</tr>
<tr>
<td>D2D UEs distribution</td>
<td>Uniform</td>
</tr>
<tr>
<td>D2D transmission power</td>
<td>-20dBm fixed ((\approx 20m \text{ range}))</td>
</tr>
<tr>
<td>Traffic load per D2D request</td>
<td>Full buffer</td>
</tr>
</tbody>
</table>
Channel bandwidth | 10MHz
---|---
Pathloss model | TS36942, urban environment
Fading | Mean: 0, std. dev.: 10dB
Thermal noise density | -174dBm/Hz
Minimum Coupling Loss | 70dB
ILI range | 1..15
SINR threshold | Configurable
$P_{out}$ threshold | 10%

**ACO parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of ants, $N$</td>
<td>Equal to the D2D requests, $D$</td>
</tr>
<tr>
<td>Evaporation rate, $r$</td>
<td>0.2</td>
</tr>
<tr>
<td>Coefficients $\alpha, \beta$</td>
<td>1, 2</td>
</tr>
<tr>
<td>Ant solution length, $L_{len}$</td>
<td>Starting value = $D$</td>
</tr>
<tr>
<td>Evaluation runs</td>
<td>2000</td>
</tr>
</tbody>
</table>

Initially, we demonstrate an example to assess the number of iterations needed for the ACO to reach a close-to-optimal solution. We allow the algorithm to run locally at the eNB for multiple evaluation runs ($x$ axis in Figure 50), and we observe the algorithm’s behavior in terms of reducing the total cost, which corresponds to the sum of $ILI_s$ across D2D connections sharing the same resources ($y$ axis in Figure 50). As it can be observed in Figure 50, the algorithm gradually approaches a stable solution, due to its self-learning nature. Starting from a reference cost of 110 units (dashed straight line), the algorithm manages to steadily reduce the total cost down to around 25 units. This reduction indicates the discovery of constantly better paths, or equivalently, the discovery of constantly better combinations of D2D requests that are more likely not to interfere with each other when sharing the same spectrum portion.

**Figure 50: Gradual convergence of ACO algorithm to a close-to-optimal solution.**

Having demonstrated the self-learning nature of the proposed algorithm, next we evaluate its performance in boosting the spatial spectrum reuse inside the cell, through the resource sharing among
concurrent D2D transmissions. We compare the performance of ACO with the performance of an optimal resource allocator, which estimates the SINR values and correlate outage probabilities by examining all possible combinations of concurrent (co-allocated) D2D transmissions, i.e., for any possible subset synthesis, and for various subset lengths. Towards this direction, we estimate the achieved spectrum reuse by both the ACO and optimal resource allocators. The spectrum reusability is measured as the amount of subsets (or colors) required to safely satisfy all scheduled requests, over the total number of D2D requests (i.e., the ratio $S/D$). In Figure 51, we plot this ratio against the density of D2D requests in the cell area under study, for various SINR thresholds, where lower target SINR thresholds imply higher tolerance to interference. Next to each plot, for the case of 2500 requests/km$^2$, the number of subsets that have been checked by each algorithm are also shown, depicting the computational consumption of each one of the approaches.

In Figure 51 we can observe that the performance of the ACO approach is comparable to that of an optimal allocator, while for lower D2D request densities the ACO seems to reach the performance of the optimal allocator. A quite noticeable result is that the number of examined subsets needed for the ACO is much lower than that for the optimal allocator. This result is a direct outcome of ACO’s self-learning nature, in the sense that knowledge accumulated by previously checked subsets is exploited towards exploring constantly better ones in the future. In other words, ACO creates a bias towards finding close-to-optimal solutions, as simulation time progresses (due to deposited pheromones), instead of considering all future solutions as equally likely.

![Figure 51: Achieved spectrum reuse by the ACO approach compared to optimal.](image)

In Figure 52, we study the effect of subset lengths to the outage probabilities observed in the network for various target SINR thresholds. The subset length is actually expressed as the percentage of D2D requests sharing the same spectrum portion over the total number of D2D requests in the cell (i.e., the ratio $L/D$).
As expected, the larger the subset length ratio, the larger the $P_{out}$. Moreover, when the target SINR threshold is high (left curve in Figure 52), the $P_{out}$ value increases rapidly for a linear increase of the aforementioned percentage. This increase is slower for lower target SINR thresholds (e.g., right curve in Figure 52), since the interference tolerance is higher.

5.2.4 Conclusions

We have presented an ACO-based approach for solving the problem of resource allocation for D2D communications in LTE-A networks via the representation of the network as a fully connected weighted graph. Using these weights, we manage to account for fine differences in interference levels among various competing D2D requests, and take resource allocation decisions accordingly. It has been demonstrated, that the ACO approach quickly converges to spectrum reuse ratios comparable to that of an optimal resource allocator that would, however, be impractical for real-time application due to its computational complexity.

Future work includes studying how the mobility of D2D users affects the decisions and convergence time of the ACO algorithm. Moreover, it would be interesting to investigate the behavior of ACO when power control mechanisms are implemented in the cell, since they would play a significant role to the powers of D2D transmitters, and consequently, to the overall interference imposed to the whole grid area.

5.3 Enabling D2D communications in LTE networks

In this section, we provide a framework for enabling explicit D2D communications in LTE networks by using already standardized features. The target is to provide a comprehensive and realistic solution for extra D2D transmissions inside an LTE cellular area, without involving the core network or changing the existing resource management policy adopted for the cellular transmissions. The proposed framework is evaluated under a moderate resource and power allocation scenario, providing the bounds of interference-free extra D2D transmissions.

5.3.1 System Model

5.3.1.1 Background

Recognizing the need for increased spatial spectrum reuse, 3GPP schedules the standardization of the D2D communications in future releases. The question that arises is whether the current research efforts on the coexistence between D2D and cellular transmissions can be applied to the already standardized...
and applied LTE network. Moving one step further, the challenge is to enable D2D communications with no performance deterioration on standard cellular communications, and without the involvement of the core network. The latter option raises the need for all the D2D procedures to be managed by the access network.

The LTE access network is called Evolved – Universal Mobile Telecommunications System (UMTS) Terrestrial Radio Access Network (E-UTRAN). The main entities of E-UTRAN are the BSs – referred to as eNBs (evolved NodeBs), and the cellular terminals – referred to as UEs (User Equipments). The communication between eNBs and UEs is organized in frames of 10ms, while each frame is divided into 10 subframes of 1ms. Referring to transmissions from and to eNBs, there are two basic categories of subframes; the downlink (DL) and the uplink (UL), respectively. During the UL subframes, UEs may request for spectrum resources from the serving eNB (if they have data to transmit), while during the DL, they interpret DL control messages to be informed about the spectrum resources that are assigned to them. The spectrum assignment for DL and UL transmissions is an eNB responsibility, and takes place on a per subframe basis. In parallel to the resource allocation procedure, a power control scheme guarantees the appropriate Signal to Noise Ratio (SNR) at the eNB and UEs during the UL and DL, respectively.

In the following section, the proposed D2D model focuses on enhancing E-UTRAN standardized procedures to allow D2D transmissions.

5.3.1.2 D2D model

The general concept of D2D communications deals with cellular-standardized UEs that have the ability to directly communicate to each other with or without the assistance of the eNB. Here, we adopt the latter case, and the eNB is able to re-allocate spectrum resources, primarily assigned to cellular users. Since we take advantage of LTE standardized features to enable the D2D communications, all UEs have the basic capability of D2D transmissions; however, hereinafter, to simplify our description, the UEs that implement our D2D scheme will be referred to as eUEs (enhanced UEs). More specifically, we adopt a model where the eNB operates as D2D controller, and as such it is responsible for the following: i) resource allocation for D2D transmissions (secondary resource allocation), ii) tuning of the D2D transmitter and receiver to the allocated resources, and iii) control of the D2D transmission power. We assume that the D2D connection establishment is handled by application layer procedures, while the proposed scheme provides eUEs with spectrum resources for D2D transmissions.

The secondary resource allocation is triggered by D2D resource requests (similar to the standardized cellular resource requests) made by eUEs. The D2D requests exclusively address resources on the UL, when the only primary (cellular) receiver in the network is the eNB, as an effort to smooth the interference problem inside the cell. The single immobile node (the eNB) is protected by a simplified D2D power control mechanism, in which the D2D transmit power is bounded. The serving eNB, which is aware of the D2D requests, re-allocates the UL cellular resources to the eUEs following a resource allocation algorithm. However, differing from the standardized resource allocation procedure, the serving eNB informs both D2D transmitter and receiver about the allocation grant, tuning them to the allocated resources. The proposed D2D model may be applied with any kind of duplex mode, either Time- or Frequency Division Duplex (TDD or FDD); nevertheless, the TDD mode is preferred due to its ability to dynamically adjust to time-varying load conditions. For instance, the UL/DL configuration “0” for Type 2 frame structure (LTE-TDD) would allow for 6 UL D2D requests per frame.

To be able to serve the D2D requests exclusively by the access network (eNBs), without any involvement whatsoever of the core network, the introduction of a new identity for each eUE is required. Note that
conventionally used destination IP addresses, or destination IMSI/S-TMSI identities (International-/Subscriber-Temporary Mobile Subscriber Identity), or other upper level explicit identities (e.g. SIP addresses) are not available locally at the eNB and thus, cannot be used for D2D paging purposes. This new identity is generated by each eUE during its initial access to the network, and any transmitting eUE has the ability to produce the D2D identity of its communication target eUE, as explained later. If an eUE wants to establish a D2D connection, the D2D identity of the target eUE is included in the D2D resource request. The serving eNB, having an one-to-one mapping between standardized and D2D identities, uses the former identities in order to inform both D2D transmitter and receiver about the resource allocation, as thoroughly described in the next section.

![Figure 53: D2D Model.](image)

The adopted D2D model can be summarized as follows (see Figure 53):

1. Each eUE produces its D2D identity (details are provided in the next section) and transmits it to the eNB during its first access to the network.
2. eUEs make D2D spectrum requests using the standard spectrum request procedure, including, however, the D2D identity of the potential D2D receiver.
3. The eNB re-allocates the UL resources for D2D transmissions, informing both D2D peers.
4. The eUE transmitter sends its data directly to the eUE receiver using the UL spectrum region allocated by the eNB, while the eUE receiver tunes to the same spectrum region to receive the transmitted data.

The LTE compatible functionalities required for enabling this D2D model are provided in the next section.

### 5.3.2 LTE Functionality for D2D Communications

The application of the adopted model to LTE networks requires enhanced functionality at the standard UEs (upgrading them to eUEs) and at the eNBs.

#### 5.3.2.1 Enhanced UE functionality

1. **D2D identity production**

   In the standardized cellular communications, the eNB uses the Cell Radio Network Temporary Identifier (C-RNTI) to uniquely identify UEs. A unique C-RNTI is assigned by the eNB to a UE during the initial random access procedure and is used for identifying the Radio Resource Control (RRC) connection and
for scheduling purposes, namely for the coding/decoding of the physical downlink control channel (PDCCH) intended for a specific UE [134].

As a result of the description of the adopted system model, the introduction of an additional D2D eUE identity is required. Hereafter, this identity will be referred to as $D2D-ID$. In contrary to the C-RNTI, the D2D-ID is produced by the eUE, using a transformation of the Mobile Subscriber Identification Number (MSIN), the 10-digit number that uniquely and globally identifies a mobile phone. All eUEs use the same algorithm for the D2D production; thus, provided that the MSIN of the target eUE is known at the eUE transmitter, the target’s D2D-ID can be faultlessly produced. The adoption of this new D2D-ID instead of directly using the MSIN is based on the need to transmit the identity of the target eUE in the UL spectrum request message by using a 16-bit MAC header portion currently reserved by the LTE standard for future use. The selection of an efficient algorithm for the MSIN transformation is out of scope.

During the Random Access (RA) procedure, the eUEs register to the network like any standard UEs. Hence, each eUE initiates a contention-based access to the network by transmitting a preamble sequence on the physical random access channel (PRACH) (Figure 54). As a result, it is supplied with a temporary random C-RNTI by the eNB via the RA Response message. Assuming that contention resolution due to potential preamble collisions is not required or is already resolved, this temporary C-RNTI will be promoted to normal C-RNTI, to be used for unique identification inside the cell, for as long as this UE stays in connected mode. The Random Access procedure is successfully completed upon the reception and the acknowledgement of the RRC Connection Setup message by the eUE. The D2D-ID is also transmitted to the serving eNB during the RA procedure. The D2D-ID is included as additional information inside the RRC Connection Request message, transmitted via the physical UL shared channel (PUSCH) (Figure 54).

### ii) D2D identity transmission

During the Random Access (RA) procedure, the eUEs register to the network like any standard UEs. Hence, each eUE initiates a contention-based access to the network by transmitting a preamble sequence on the physical random access channel (PRACH) (Figure 54). As a result, it is supplied with a temporary random C-RNTI by the eNB via the RA Response message. Assuming that contention resolution due to potential preamble collisions is not required or is already resolved, this temporary C-RNTI will be promoted to normal C-RNTI, to be used for unique identification inside the cell, for as long as this UE stays in connected mode. The Random Access procedure is successfully completed upon the reception and the acknowledgement of the RRC Connection Setup message by the eUE. The D2D-ID is also transmitted to the serving eNB during the RA procedure. The D2D-ID is included as additional information inside the RRC Connection Request message, transmitted via the physical UL shared channel (PUSCH) (Figure 54).

---

**Figure 54: D2D enhancements in standard LTE signalling.**
### iii) D2D connection request

Let eUE₁ want to establish a D2D connection to eUE₂. Assume that both eUEs have already submitted their D2D-IDs using the D2D-ID information element in their RRC connection request message. Normally, when a UE has data to transmit, the Buffer Status Report (BSR) procedure is initiated [152]. According to this procedure, a Regular BSR informs the serving eNB via the PUSCH about the amount of data pending for transmission in its UL buffers. Note that if no BSR is already allocated (perhaps because no other transmissions are already initiated), a single-bit Scheduling Request (SR) on the physical UL control channel (PUCCH) precedes the BSR request [152].

![Figure 55: Enhanced MAC PDU.](image)

In addition to the standard information that any UE includes in the BSR request, the eUE₁ produces the target UE’s (eUE₂) D2D-ID and adds it to this request (Figure 54). To this end, an unused 16-bit long MAC Control Element is used. It is the presence of this extra Control Element that declares to the eNB that a D2D request is initiated instead of a cellular one. This element utilizes space currently reserved for future use and it is indexed in the MAC Protocol Data Unit (PDU) subheader by the Logical Channel ID (LCID) value equal to 11000. The new element is called D2D Receiver ID and is appended to the existing LCID values, such as the common control channel (CCCH), the C-RNTI and the Padding [152]. The MAC PDU structure that includes D2D requests is shown in Figure 55. In this figure, the extra MAC sub-header for D2D request is depicted as the last subheader of the MAC header.

#### 5.3.2.2 Enhanced eNB functionality

**i) D2D identities’ storage at eNBs**

During a standard initial (e)UE attachment to the network, the eNB assigns a unique C-RNTI to the requesting device. This unique identity is later used by the (e)UE to address itself to the eNB, and in fact it is also included in the RRC Connection Request message. Additionally, each (e)UE isolates the resources that belong to it by performing blind decoding to the UL/DL allocation grant using its own C-RNTI.

According to the proposed scheme, when an eUE requires resources for a D2D transmission, it includes the D2D-ID of the D2D destination in the request message (D2D Receiver ID). The eNB keeps an one-to-one mapping between C-RNTIs and D2D-IDs, created during the initial network access of each eUE, as previously explained. Consequently, upon the reception of this request, the eNB locates the C-RNTIs of the requesting and the destination eUEs in this mapping table using the respective D2D-IDs. Following that, it uses the corresponding C-RNTIs to encode the allocation messages for the eUE transmitter and receiver.

**ii) Resource allocation**

The eNB uses a resource allocation algorithm in order to perform a secondary UL grant to the requesting eUEs through the PDCCH. As mentioned before, the secondary allocation grant is also transmitted to the
target eUEs (D2D receivers) to inform them about the appropriate spectrum region in which they will receive the D2D data. We refer to the resource allocation to the D2D receiver as D2D RX grant to differentiate it from the D2D TX grant of the D2D transmitter, even though both refer to the same physical resources (Figure 54).

An adopted secondary resource allocation scheme is a key factor for guaranteeing fair, reliable, and interference-free spectrum sharing between cellular and D2D communications, as well as among D2D communications themselves. In the general case, the secondary resource allocation scheme can be independent of the cellular/primary one. However, as shown in [129], the primary resource allocation algorithm includes important information that can be used for the design of an interference-aware secondary allocator. For example, if a D2D pair uses UL resources that are allocated to a single UE, the interference perceived by the eNB is strongly dependent on the locations of the UE and eUE transmitters. On the other hand, the use of resources that are allocated to more than one UEs could reduce the impact of the (e)UEs’ locations on the interference imposed due to the geographical multiuser diversity.

In the best case scenario in terms of spectrum utilization, no resources are instantaneously assigned for cellular transmissions (e.g., due to low cellular traffic), and thus all available spectrum resources will be shared solely among D2D pairs, potentially leading to very high spatial spectrum reuse.

Here, we assume a simple secondary resource allocator applicable to the LTE physical layer, where each D2D link uses resources primarily assigned to a single UE, independently of the standard allocation scheme. To achieve this independency, the minimum standardized allocation unit can be used as the fixed allocation unit for a D2D communication. In this way, the secondary allocator indirectly avoids allocating resources of different UEs to a single D2D pair, significantly limiting the D2D outage probability. In LTE networks, the minimum allocation unit is defined by a Physical Resource Block (PRB) as explained in [134].

### iii) Power control

Power control is the key factor for interference protection of the serving eNB during the UL. An efficient and dynamic approach is described in [126], where the view of the standard power control mechanism is reversed: instead of resulting in the lowest power needed for a target SNR, the maximum interference-free transmit power is achieved. However, as shown in [153], the use of the maximum interference-free power can lead to a low spatial spectrum reuse factor, and in the best case we can reuse the spectrum up to 6 times in a cell where the eNB is located at the center of it. In line with the concept of short range D2D communications, and with respect to the requirement of not changing the primary power control scheme, we use fixed power for the D2D communications. Taking into account the LTE standard, we assume that for a short-range D2D communication (up to 50 meters), the eUEs use a fixed power level that guarantees satisfactory connection in that range. This choice allows us to examine the upper bound on the number of concurrent short-range D2D communications in a cell.

### 5.3.3 Evaluation Results

For the evaluation of the proposed scheme, we made changes to the system level simulator proposed in [150] in order to support D2D communications. We focus on a single UL cellular transmission which coexists with multiple D2D connections. The basic parameters of the simulation are shown in Table 12.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
</table>

Table 12: Basic simulation parameters.
In Figure 56, we examine the maximum number of D2D devices that can transmit concurrently with a UE transmitter located at a distance of half the cell radius from the eNB. The UE, through the power control procedure, transmits with the minimum power that guarantees a target SINR threshold at the eNB. D2D transmitters use fixed power that equals to -19 dBm, a power level that can guarantee an acceptable SINR level at a small distance (up to 50m). As it was expected, the more the D2D transmitters the more the SINR degradation at the eNB. However, an interesting result extracted by Figure 56, is that for D2D users located at a distance more than ~35% of the cell radius from the eNB, the SINR distortion is negligible and practically independent of the number of D2D transmitters. This result validates that the D2D connectivity is a suitable candidate for improving the spatial spectrum utilization, especially at areas closer to cell edges. From the opposite perspective, the protection of the D2D connections from the UE and the other D2D transmissions is also very important. In this direction, a comprehensive result is given in Figure 57.

![Figure 56: SINR at eNB for different number and locations of D2D transmitters.](image)

In Figure 57, we calculate the percentage of the cell area where potential D2D receivers are instantaneously not interfered by the cellular UE’s transmission. This percentage is actually a measure of
the maximum spatial spectrum reuse ratio. As shown in this figure, the dependency of the spectrum reusability on the distance of the cellular UE from the eNB is higher as the target SINR at the eNB increases. The higher the required quality for a cellular communication, the higher the required SINR at the eNB, and, thus, the higher the transmit power of the UE, something that will inevitably reduce the allowed D2D reuse. Moreover, since the UE transmitter uses an omni-directional antenna, its signal best covers the cell area when it is located at a distance of approximately half the radius from the eNB, suppressing the possibility for parallel non-interfering D2D transmissions, which justifies the concave shape of the curves. In the case of multiple D2D transmitters reusing the same resources, this ratio is expected to decrease, because neighboring D2D transmissions will cause interference to each other. Generally, during each UL cellular transmission, around 40% to 100% of the cell area seems to be available for extra D2D transmissions.

![Figure 57: Spatial spectrum reuse ratio in whole cell for different UE locations.](image)

5.3.4 Conclusions and Potential Applications

We have provided a framework for enabling D2D communications in LTE networks by using already standardized features. We described a comprehensive solution for extra D2D communications without involving the core network or changing the resource management policy adopted for the cellular transmissions. Simulation results have shown that even a single cellular UL transmission of an LTE network can be reused by multiple D2D connections with respect to the SINR at the eNB, while the opportunities for spatial spectrum reuse span a satisfactory fraction of the cell area.

The combination of the proposed framework with efficient secondary resource allocation and power control mechanisms can facilitate several new applications. For example, applications for emergency relief can benefit from D2D communications in the case that the core network is down but the access network (eNB) is still functional. Also, applications that make daily life easier can utilize the proposed framework, as for instance in the case that two friends are sitting next to each other and want to share some content. Finally, users subscribed to specific advertising services can receive notifications via D2D connections when they are located close to relevant offers.

5.4 Future work

Future work includes the studying of D2D in combination with UL/DL Decoupling (DUDe). This study will be done in two directions. First, it would be interesting to investigate how the UL/DL decoupling of standard cellular users will have an effect on the imposed interference and hence, resource efficiency of
D2D users, when spectrum is shared. Next, we could consider that the requirements of the D2D requests trigger the DUDe selection of the cellular users, so as to maximize the spectrum reuse in the cell. Second, we would like to investigate the possibility of conducting some kind of UL/DL decoupling for the D2D users themselves. To do this, we first have to define what DUDe means for D2D users, how this can be implemented, and which are the potential benefits of such an approach.

6. Conclusion

This document showed the advances done by the consortium about the algorithms for SON and cognition in LTE-A communication system, where the work from each ESR has been presented in detail. Each ESR tackled a different topic within the field and provided excellent contributions that paved their way into technical publications. Five topics have been considered: the D2D communication, the handovers optimization, the traffic offloading, the elastic resource sharing over TDD/FDD systems and the resource sharing optimization; where each chapter presented the achievements within each selected topic, and a review of the state of the art in first introduced in order to realize the current status
on the science in the research arena, and to show the advantages of our proposals in comparison to the SoA.
References


[56] Sha Hua; Xuejun Zhuo; Panwar, S.S., "A truthful auction based incentive framework for femtocell access," Wireless Communications and Networking Conference (WCNC), 2013 IEEE , pp.2271,2276, 7-10 April 2013


[67] Li, Yanwei; Zhou, Zhenyu; Nguyen, Nam; Sato, Takuro, "Game Theory Based Hybrid Access for Macrocell-Edge Users in a Macro-Femto Network," Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th , pp.1.5, 2-5 June 2013


[69] Yanjiao Chen; Jin Zhang; Qian Zhang, "Incentive mechanism for hybrid access in femtocell network with traffic uncertainty," Communications (ICC), 2013 IEEE International Conference on, pp.6333,6337, 9-13 June 2013


Yangming Zhao; Sheng Wang; Sizhong Xu; Xiong Wang; Xiujiao Gao; Chunming Qiao, "Load balance vs energy efficiency in traffic engineering: A game Theoretical Perspective," INFOCOM, 2013 Proceedings IEEE , pp.530,534, 14-19 April 2013


P.R. Winters: "Forecasting Sales by Exponentially Weighted Moving Averages, Management Science, vol. 6, pp. 324-342, 1960


Panagiotis Trakas, Ferran Adelantado, Christos Verikoukis, “A novel learning mechanism for traffic offloading with Small Cell as a Service,”IEEE International Communications Conference on, 8-12 June (accepted) 2015

NGMN alliance, "5G White Paper", Feb 2015


4G Americas Mobile Broadband Evolution Rel-10, Rel-11 and Beyond-HSPA+, SAE/LTE and LTE-Advanced White Paper

4G Americas Mobile Broadband Evolution Rel-11 & Rel-12 Progress White Paper


Recommended Practices for Multi-vendor SON Deployment Deliverable D2 by NGMN Alliance

Telecommunication management; Self-Organizing Networks (SON) Policy Network Resource Model (NRM) Integration Reference Point (IRP); Information Service (IS), (Release 11)

Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements (3GPP TS 32.500 (Release 11))

3GPP TS 36.423 V11.3.0, X2 application protocol (X2AP), (Release 11)
[103] 3GPP TR 32.860 V0.6.0, Study on Enhancements of OAM aspects of distributed Self-Organizing Networks (SON) functions (Release 12)


