Initial Training Network

CROSSFIRE

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

FP7 Contract Number: 317126

WP2 – Network Virtualization in LTE-A Networks

D2.2

Network Virtualization Component Level Algorithms for Mobility Management and RRM
<table>
<thead>
<tr>
<th>Contractual Date of Delivery:</th>
<th>February 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Date of Delivery:</td>
<td>February 2016</td>
</tr>
<tr>
<td>Responsible Beneficiary:</td>
<td>NEC</td>
</tr>
<tr>
<td>Contributing Beneficiaries:</td>
<td>UoC, UoA, NEC, KCL</td>
</tr>
<tr>
<td>Security:</td>
<td>Public</td>
</tr>
<tr>
<td>Nature:</td>
<td>Report</td>
</tr>
<tr>
<td>Version:</td>
<td>1.0</td>
</tr>
</tbody>
</table>
### Document Information

**Version Date:** 20/02/2016  
**Total Number of Pages:**

### Authors

<table>
<thead>
<tr>
<th>Name</th>
<th>Organisation</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Georgia Tseliou</td>
<td>UoC</td>
<td><a href="mailto:gtseliou@uoc.edu">gtseliou@uoc.edu</a></td>
</tr>
<tr>
<td>Salvatore Costanzo</td>
<td>UoA</td>
<td><a href="mailto:scostanzo@di.uoa.gr">scostanzo@di.uoa.gr</a></td>
</tr>
<tr>
<td>Rudraksh Shrivastava</td>
<td>NEC</td>
<td><a href="mailto:rudraksh.shrivastava@neclab.eu">rudraksh.shrivastava@neclab.eu</a></td>
</tr>
<tr>
<td>Giorgos Chochlidakis</td>
<td>KCL</td>
<td><a href="mailto:giorgos.chochlidakis@kcl.ac.uk">giorgos.chochlidakis@kcl.ac.uk</a></td>
</tr>
</tbody>
</table>

### Document History

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Modification</th>
<th>Contact Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>20/02/15</td>
<td>First ToC</td>
<td>K. Samdanis</td>
</tr>
</tbody>
</table>
Executive Summary

This deliverable presents an overview the Work Package 2 (WP2) of CROSSFIRE (unCooRdinated netwOrk StrategieS for enhanced interFerence, mobIilty, radio Resource, and Energy saving management in LTE-Advanced) project. The deliverable continues the work from the first document that concentrated on the state-of-the-art and explores more details considering network architectures, algorithms and mechanisms that can address network virtualization in a wireless environment concentrating on LTE-A systems. In particular, it analyses two thematic aspects of network virtualization in LTE-A systems considering radio and network resource sharing as well as network function virtualization. For the radio and network sharing this deliverable explores (i) a 3GPP compliant approach considering the Multi-Operator Core Network and Gateway Core Network architectures for small cell heterogeneous environments and (ii) an open eNB approach based on the ONF OpenFlow technology, which enables operators to acquire base station resources on-demand. Considering the network function virtualization, this deliverable explores (i) radio access network virtualization concentrating on TD-LTE systems considering radio resource virtualization and virtual cells and (ii) core network virtualization concentrating distributed mobility management and mobility aware virtual network embedding.
Table of Contents

Executive Summary........................................................................................................iv

1. Introduction...............................................................................................................10

2. RAN and Network Sharing (UoC)........................................................................12
   2.1 Resources Negotiation in Small Cell Deployments ......................................13
       2.1.1 Introduction ..............................................................................................13
       2.1.2 Related Work on the proposed solution ................................................14
       2.1.3 Proposed Scheme .....................................................................................16
           2.1.3.1 Network Deployment ........................................................................16
           2.1.3.2 Resource Management in LTE-A ..................................................16
           2.1.3.3 Resources Negotiation for Network Virtualization Algorithm .......17
       2.1.4 Analytical Formulation ..............................................................................18
           2.1.4.1 A. System Model and Assumptions ..................................................18
           2.1.4.2 Problem Setup ..................................................................................19
           2.1.4.3 Formulas of Aggregate Throughput ................................................20
           2.1.4.4 Analysis .............................................................................................20
       2.1.5 Performance Evaluation ............................................................................22
       2.1.6 Conclusions ...............................................................................................25

   2.2 Scalable RAN Virtualization in Multi-Tenant LTE-A HetNets ..................26
       2.2.1 Introduction ..............................................................................................26
       2.2.2 State Of The Art and Contribution ...........................................................27
       2.2.3 RENEV in a Heterogeneous Scenario ....................................................30
           2.2.3.1 Scenariounderconsideration .............................................................30
           2.2.3.2 Radio Resource Management Functions .......................................31
           2.2.3.3 Proposed Algorithm: RENEV for HetNet LTE-A deployments .....32
       2.2.4 Signaling Design Considerations ...............................................................36
           2.2.4.1 Detectionphasesignaling ..................................................................37
           2.2.4.2 Transfer phase signaling ..................................................................38
       2.2.5 Throughput Analysis ..................................................................................39
           2.2.5.1 System Model ....................................................................................39
           2.2.5.2 General Throughput Formulation ....................................................39
       2.2.6 Additional Signaling Overhead Analysis ....................................................43
2.2.7 Performance Evaluation ......................................................... 47
  2.2.7.1 Simulation Scenario and Parameters ..................................... 47
  2.2.7.2 Network Performance .................................................. 49
  2.2.7.3 User’s Throughput ..................................................... 51
  2.2.7.4 Signaling Overhead .................................................... 52
2.2.8 Concluding Remarks ....................................................... 54

2.3 Appendices .............................................................................. 54
  2.3.1.1 Appendix A: MCS Selection Probability .............................. 55
  2.3.1.2 Appendix B: Derivation Of Equation (6) ............................... 56
  2.3.1.3 Appendix C: Set Of Feasible Future States ......................... 59
  2.3.1.4 Appendix D: Derivation of $P(Q = q | N, M)$ ....................... 60

3. Open eNB (UoA) ........................................................................ 61
  3.1 Proposed Architecture .......................................................... 63
    3.1.1 OpeNB ........................................................................... 64
    3.1.2 Main Controller ............................................................. 65
    3.1.3 L3 tunnel ................................................................. 66
  3.2 OpeNB Sharing Procedure .................................................... 67
    3.2.1 NV-Triggering Phase .................................................... 69
    3.2.2 NV Decision ............................................................... 71
    3.2.3 SLA Monitor ............................................................... 71
    3.2.4 NV Admission Control ................................................ 72
  3.3 Simulation Analysis ............................................................... 72
  3.4 CONCLUSIONS ..................................................................... 81

4. Programmable RAN (NEC) ......................................................... 82
  4.1 Concept Description/Programmable RAN for TD-LTE .................. 82
  4.2 Related Work ......................................................................... 84
  4.4 Virtual Cell Concept ............................................................. 86
  4.5 SDN-based Network Management Architecture .......................... 87
  4.6 SINR Analysis and Effective Capacity Allocation for Virtual Cells .... 89
  4.7 Elastic Resource Management Algorithm for TD-LTE ............... 92
  4.8 System Simulation and Result Analysis’ .................................. 97
  4.9 Conclusion ............................................................................ 103
5. Core Network Virtualization and Mobility Management (KCL)......104
   5.1 Introduction........................................................................................................104
   5.2 Hybrid DMM for Next-Generation Wireless Networks.................................104
      5.2.1 Introduction ..................................................................................................105
      5.2.2 Selected Related Works .............................................................................107
      5.2.3 System Model .............................................................................................108
      5.2.3.1 Centralized mobility management scheme ............................................108
      5.2.3.2 Distributed Mobility Management .........................................................112
      5.2.3.3 Hybrid Distributed Mobility Management .............................................113
      5.2.4 Numerical Evaluation .................................................................................114
      5.2.5 Conclusions ...............................................................................................118
   5.3 Mobility Aware Virtual Network Embedding.................................................118
      5.3.1 Introduction .................................................................................................119
      5.3.2 Previous related work ................................................................................121
      5.3.3 System Model Description .........................................................................123
      5.3.3.1 Implementation of virtual network embedding algorithms ..................123
      5.3.3.2 Mathematical programming setting .......................................................124
      5.3.3.3 Greedy Heuristic Algorithm ..................................................................128
      5.3.4 Numerical Investigations ...........................................................................129
      5.3.5 Conclusions ...............................................................................................134
   5.4 Future Steps.......................................................................................................134

6. Conclusions ..........................................................................................................135

List of Acronyms .......................................................................................................137

References ...............................................................................................................138
Figure 1. Aggregate System Throughput for different number of offered loads........23
Figure 2. Percentage of transferred Resource Blocks...........................................25
Figure 3. Heterogeneous two-tier deployment with SCs deployed in clusters...........30
Figure 4. RENEV for a Heterogeneous Deployment..............................................33
Figure 5. Call Flow of the messages for (a) UE Attachment and RENEV and (b) UE
Context Exchange and RENEV.................................................................37
Figure 6. Aggregate System Throughput for different number of Offered Loads.......49
Figure 7. (a) Percentage of transferred RBs by each tier. (b) Traffic Served by each
tier...........................................................................................................50
Figure 8. CDFofuserThroughputfor(a)42Mbps,(b)66Mbpsand(c)78Mbps
OfferedLoad.............................................................................................52
Figure 9. Percentage of successful requests for different number of SCs per
cluster.........................................................................................................53
Figure 10. NumberofexchangedX2messagesperSC...............................................54
Figure 11: Proposed Multi-tenant RAN Architecture..............................................64
Figure 12: OpeNB Sharing Signaling Flow..........................................................68
Figure 13: Average DL Goodput for UEs registered to Operator A......................75
Figure 14: Average DL Goodput for UEs registered to Operator B......................76
Figure 15: Average DL SINR for UEs registered to Operator A............................77
Figure 16: Average DL SINR for UEs registered to Operator B............................78
Figure 17: Average packet delay for UEs registered to Operator A......................78
Figure 18: Average packet delay: UEs registered to Operator B...........................79
Figure 19: HO signaling rate for the UEs registered in Operator A..........................80
Figure 20: A Simple Example of the Virtual Cell Concept.................................86
Figure 21: SDN-Based Network Architecture.....................................................88
Figure 22: Throughput Downlink bits/second.....................................................98
Figure 23: Throughput Uplink bits/second..........................................................99
Figure 24: Mean and Low SINR Users’ Throughput Normalized to the Virtual Cell
Configuration in Downlink...........................................................................99
Figure 25: Mean and Low SINR Users’ Throughput Normalized to the Virtual Cell
Configuration in Uplink................................................................................100
Figure 26: Delay CDF Downlink (ms)................................................................101
Figure 27: Delay CDF Uplink (ms).....................................................................101
Figure 28: Mean and Low SINR Users’ Delay Normalized to the Virtual Cell
Configuration in Downlink............................................................................102
Figure 29: Mean and Low SINR Users’ Delay Normalized to the Virtual Cell Configuration in Uplink .................................................................102

Figure 30 (a) Centralized scheme, (b) DMM scheme and (c) HDMM scheme ........108

Figure 31 Total cost for different topologies ..................................................116

Figure 32 Total cost for different mobility scenarios ......................................116

Figure 33 Total cost for each AR domain ........................................................117

Figure 34 Performance of DMM and HDMM for different mobility scenarios ...117

Figure 35 (a) Mobility agnostic and (b) mobility aware virtual network embedding algorithms ....................................................................................121

Figure 36 Dense tree topology modelling a substrate network with 31 nodes ....131

Figure 37 Sparse tree topology modelling a substrate network with 31 nodes ....131

Figure 38 Impact of users mobility effect on the total routing cost for mobility aware and unaware VNE algorithms .................................................................132

Figure 39 Acceptance rate vs. number of different alternative paths ...............132

Figure 40 Acceptance rate vs. mean node capacity value ...............................133

Figure 41 The impact of the substrate topology .............................................133
1. Introduction

The CROSSFIRE (unCooRdinated netwOrk StrategieS for enhanced interFerence, mobility, radio Resource, and Energy saving management in LTE-Advanced) project investigates a number of evolving issues and challenges regarding LTE-Advanced (LTE-A) networks. In particular, it explores network architecture, operations and management aspects as well as mechanisms for enhancing the user performance considering interference mitigation, congestion management and Quality-of-Experience (QoE). This deliverable contains network architectures, algorithms, and mechanisms for supporting network virtualization in LTE-A systems. The focus of this deliverable is to present a first round of the proposed novel mechanisms that were the outcome of CROSSFIRE, presenting also performance evaluation and analytical results, which quantify and demonstrate how CROSSFIRE can enhance the state of the art.

In particular this deliverable contains architecture and mechanisms for addressing radio and network sharing for LTE-A based on the current 3GPP architecture. To begin with, a novel algorithm for resources virtualization called Resources nEgotiation for Network Virtualization (RENEV) is introduced. RENEV realizes the process of abstracting resources, by customizing them in isolation among different BSs that have specific requirements due to traffic non-uniformities by operators. In this solution, distributed BSs, located in a shared RAN among operators, cooperate to reallocate radio resources on a traffic need basis. RENEV is based on the concept of physical resources transfer, defined as the possibility of reconfiguring the Orthogonal Frequency Division Multiple Access (OFDMA)-based medium of two BSs, to allow a BS to use a set of subcarriers initially allocated to another BS. This solution can be applied into network sharing configurations defined by 3GPP (i.e., Gateway Core Network and Multi-Operator Core Network) with reasonable signaling cost after identifying the limitations into the architecture. A simulation evaluation and an analytical study demonstrate that RENEV can enhance the performance of some state of the art virtualization proposals up to 52.66%.

In addition, a complementing architecture, mechanisms and network protocol is introduced based on ONF OpenFlow for enabling radio and network sharing considering Open eNBs, i.e. base stations that can be shared on-demand by different operators. The proposed architecture aims at virtually increasing the base station density of a mobile network operator by enabling on-the-fly leasing of the physical networking infrastructure from
other mobile network operators. The aforementioned process is performed based on a-priori service level agreements (SLAs) established by the mobile network operators involved, while it takes place under certain conditions related to the network status of the two operators, e.g. when additional base stations are needed for traffic offloading, or when link quality improvements can be achieved. In this work, triggering and network virtualization decision algorithms have been presented to effectively employ the proposed dynamic sharing of base stations based on predefined SLAs. A simulation study showed that the proposed architecture can result in up to 17% of improvement in terms of throughput and in substantial reduction (up to 50% gain) of the end-to-end application-layer delay at the users of the home operator, i.e. the sending operator, without significantly affecting the performance of the users at the host operator, i.e. the operator that shares its infrastructure.

For virtualizing the radio access network this deliverable concentrates on TD-LTE introducing the concept of virtual cell, in where sub-frame resources from neighboring eNBs can be combined to introduce a virtual customized frame addressing the needs of residing user’s applications. Mechanisms and algorithms for efficiently managing the resources of a TD-LTE network in a flexible manner have been proposed, enabling (i) dynamic frame alternation at each eNB and (ii) forming virtual cells, which allow diverse resource utilization to users residing within regions that can utilize resources from multiple eNBs. Our approach leverages the benefits of the Software Defined Networks (SDN) paradigm for monitoring network resource utilization and allowing applications or services to request resources. The resources requested by the applications or services can be allocated on-demand by adjusting the TDD frames in different geographical regions considering also the creation of virtual cells in overlapping regions that can serve best the residing users. The performance evaluation results indicate significant performance gains between 30-35% considering both UL/DL directions respectively as compared to the conventional state of the art mechanisms.

Considering the network virtualization for the core network this deliverable introduces a hybrid distributed mobility management approach and a mobility aware virtual network embedding solution for LTE-A systems and beyond. The performance benefit of the proposed mobility management scheme can be up to 17% in comparison to so far proposed distributed mobility management schemes considered in IETF. Moreover, the virtual network embedding solution that explicitly accounts the users’ mobility effect is proposed and evaluated outperforming existing algorithms from the literature in terms of routing cost (up to
47% gain) as well as admission rate being successful in scenarios where other solutions fail to serve.

The remaining of the deliverable is organized as follows. Section 2 presents the 3GPP compliant radio and network sharing mechanisms for heterogeneous scenarios, while section 3 the Open eNB architecture and protocol. Section 4 details the concept of virtual cells for TD-LTE and elaborates the corresponding network management architecture, which is based-on SDN. Section 5 documents the hybrid distributed mobility management solution and the mobility aware virtual network embedding. Finally, section 6 provides the conclusions.

2. **RAN and Network Sharing (UoC)**

This section presents proposed solutions for Radio Access Network (RAN) virtualization and Network sharing.

Radio Access Network (RAN) virtualization is a promising commercial solution where multiple service providers can share underlying radio physical resources and dynamically compose heterogeneous virtual networks that coexist in isolation within the same physical infrastructure. Based on the expected future requirements, section 2.1 proposes a solution named Resources nEGotiation for NEtwork Virtualization (RENEV), which can be applied in Long Term Evolution-Advanced (LTE-A) environments, consisting of numerous small cells. This algorithm aims at achieving an efficient mapping of radio virtual network elements onto the radio resources of the existing physical network, utilizing the concept of radio resource transfer. It establishes a common virtualized control layer by handling the resources in an holistic way. The proposed solution achieves significant gains in terms of system throughput and its performance is evaluated by means of analytical model, as well as simulation results.

In addition section 2.2 modifies the RENEV algorithm, suitable for application in Heterogeneous Networks (HetNets) in LTE-A environments, consisting of a macro evolved NodeB (eNB) overlaid with small cells. By exploiting Radio Resource Management (RRM) principles, RENEV achieves slicing and on demand delivery of resources. Leveraging the multi-tenancy approach, radio resources are transferred in terms of physical radio Resource Blocks (RBs) among multiple heterogeneous base stations (i.e., tenants), interconnected via the X2 interface. The main target is to deal with traffic variations in geographical dimension. All signaling design considerations under the current Third Generation Partnership Project
(3GPP) LTE-A architecture are also investigated. Both analytical studies and simulation experiments are conducted to evaluate RENEV in terms of network’s throughput as well as its additional signaling overhead. Moreover we show that RENEV can be applied on top of already proposed schemes for RAN virtualization to improve their performance. The results indicate that significant merits are achieved both from network’s and users’ perspective as well as that it is a scalable solution for different number of small cells.

2.1 Resources Negotiation in Small Cell Deployments

2.1.1 Introduction

The management and provision of services of current networks, cannot match the more demanding and fast changing requirements imposed by end-user applications. Due to the existence of multiple stakeholders and the need of reducing deployment costs, network virtualization has been proposed as a promising technique to overcome the ossification of the current Internet. Coupled with an effective and efficient approach to manage virtualized resources, it is expected to be the central element of future architectures.

Today’s cellular networks have relatively limited support for virtualization. LTE-A supports the isolation of different enterprise customers’ traffic into Virtual Private Networks (VPN) using traditional Border Gateway Protocol (BGP) / Multi-protocol Label Switching (MPLS) VPN technologies. However, one of the main open issues is that LTE-A does not allow different carriers to share the infrastructure to offer a complete virtual LTE-A solution to their customers. In addition, current sharing solutions have limitations in terms of separating control and data planes among operators, accommodating different requirements per operator, and adapting updated requirements. Consequently, the necessity for sharing the existing and future infrastructure among several operators and service providers in a seamless and isolated way is imminent.

Moreover, the necessity for network virtualization becomes even more important in dense architectures. The latest release of LTE-A has been especially focused on small cells, i.e. Home Evolved Universal Terrestrial Radio Access (E-UTRA) NodeBs (HeNBs) [1] LTE-A offers the opportunity of creating very dense architectures, consisting of multiple tiers but the network sharing architecture proposed by 3GPP allows different core network operators to connect to a shared RAN without exploiting this heterogeneity [13]. In the end nowadays, a single network market with operators competing at the service layer is considered a commercially viable model.
The motivation for this work is twofold: we want to utilize the concept of sharing resources in terms of physical RBs in on-demand and self-service way, to achieve the principles of RAN virtualization in LTE-A, and in parallel take advantage of the newly introduced concept of small cells to offload the available data traffic. The contributions can be summarized into the following points:

- A solution based on resources transfer is presented: the Resources nEgotiation for NEtwork Virtualization (RENEV) algorithm. It is a solution of mapping virtual path requests to substrate physical topology. It is a step towards RRM in the context of network virtualization without altering the nodes of the existing architecture.
- We focus on how to utilize the fractional radio resources spread around on multiple physical small cells using the concept of physical resource transfer. This concept is an efficient way to redistribute the available resources and help small cells to serve seamlessly the maximum possible number of users.

2.1.2 Related Work on the proposed solution

Network virtualization area in cellular networks field requires more attention. Although some results have recently been published, there is still a great need for more comprehensive ones addressing relevant issues so as to support the emerging content-rich end-user services in a cost-effective way. Network virtualization, as a concept, can be applied in several sections of the network; also in the case of cellular networks two options for its application arise: the Evolved Packet Core (EPC) Network and the RAN. We are going to focus on the RAN side of LTE-A that makes possible the easy creation and management of virtual networks, opening up a range of new business models.

Spectrum sharing is a key technique in LTE-A RAN virtualization; it can be used at the radio interface to adapt to the traffic load variation of different virtual networks. The term eNodeB virtualization refers to the case where multiple virtual networks share the spectrum of the same physical eNodeB. In [21] a scheduling framework of the available spectrum between different eNodeBs belonging in different operators is proposed. A controlling entity called hypervisor is also included to make use of a priori knowledge in order to schedule the RBs.

A feature called Network Virtualization Substrate (NVS) has been proposed by [24] and [25], for managing and sharing the radio spectrum and eNodeB processing resources. In this work, a slice scheduler which works in conjunction with the Medium Access...
Control (MAC) scheduler monitors the amount of resources that the MAC scheduler assigns to each slice of a Macro eNodeB; so it dynamically adjusts the bearer priorities in the MAC scheduler to maintain the required resource allocation for each operator.

Dynamic resources' slicing is another category of solutions based on the concept of RAN virtualization. The authors of [23] have proposed CellSlice; a dynamic framework to achieve active RAN sharing by remotely controlling the scheduling decisions, ensuring that each entity receives its share of the wireless resources. This idea does not require the modification of the Base Station (BS) schedulers but it constrains the BS scheduling decisions from a remote gateway.

In [18] the authors present a software defined cellular network architecture that allows controller applications to express policies based on the attributes of subscribers, rather than network addresses and locations, enables real-time, fine-grained control via a local agent on each switch, and extends switches to support features like deep packet inspection and header compression to meet the needs of cellular data services. Finally, in study [19] that completes the previous work, the same authors add to the proposed system an entity called CellSDN controller. Its design has as target to separate traffic management from the low-level mechanisms for installing rules and minimizing data-plane state.

Obviously, researchers have a tendency to focus on slicing the available spectrum instead of proposing solutions where the spectrum is delivered on-demand to the participating operators. Since LTE-A is an Orthogonal Frequency Division (OFDM) based system, the whole system bandwidth is available for an eNodeB. However, not all the subcarriers are used simultaneously in a specific set of cells, i.e. according to Inter-Ce Interference Coordination (ICIC) techniques each subcarrier is never allocated to more than one BS simultaneously. Accordingly, an efficient use of the available radio resources can be achieved if a proper coordination / negotiation of the resources is carried out among the BSs. Moreover, to the best of our knowledge, there is no previous work that takes into account the concept of resources negotiation by exploiting the dense LTE-A environment. RENEV presents a practical solution based on network virtualization to implement such resources negotiation. As the minimum resources unit defined in LTE-A is the RB (a 0.5msec time slot with 180KHz bandwidth), our work proposes the negotiation of the resources among BSs on RB basis.
2.1.3 Proposed Scheme

2.1.3.1 Network Deployment

In principle, a wireless cellular network deployment has two main objectives. Firstly it should guarantee coverage continuity by minimizing the outage probability and interference levels and secondly it should be able to reduce the cost of the deployment itself. Although cellular network deployment has been traditionally modeled with regular hexagonal shaped cells, this is no longer valid in current and future networks. The inclusion of small cells, e.g. HeNBs, along with the irregular traffic distribution poses new challenges in the management of the radio resources. In such a context, the underlying considered network, is a residential region composed of an eNodeB and a number of HeNBs located close to each other in random positions [13]. Throughout the rest of this work, we refer with the general term BS to all the RAN nodes of the system, both eNodeBs and HeNBs.

When a user is served by a certain HeNB, this latter is called serving HeNB of the user. Here, an open access small cell network is considered and we assume that the downlink transmitted power is fixed and the same for all the HeNBs. In particular such consideration is efficient in today’s 3GPP LTE-A where the same amount of power is transmitted on all RBs and there is no or very limited power control in the downlink.

2.1.3.2 Resource Management in LTE-A

As radio resources are scarce and at the same time the number of users and their demands increase, their management becomes a crucial point in the performance of any wireless network. In LTE-A, small cells are involved in distributed control protocols in order to manage various procedures such as handoff, interference etc. They also handle admission control and radio resource allocation. Radio resource is an inherently shared characteristic among RAN nodes, so the lack of central control leads to several difficulties in optimizing the radio access related tasks [2].

Our proposal focuses on the Access Stratum (AS) of the LTE-A. The control plane of a RAN node can be logically splitted in two entities: the baseband and the network module. The baseband module is responsible for bearer configuration with the users via the Radio Resource Control (RRC) protocol whereas the network module connects the BS with the EPC network. RRC, is a Layer 3 AS protocol of the control plane layer that handles the User Equipment (UE) management and controls Layer 2 and Layer 1 parameters, as well as UE - eNodeB Signalling by transferring common dedicated information. Furthermore, the latest
3GPP specification of LTE-A, supports the direct control and data information exchange between HeNBs via the definition of the point-to-point, logical X2 interface. This occurs independently of whether any of the involved HeNBs is connected to an intermediate node or not [13].

RRM is an eNodeB application level set of functions, ensuring the efficient use of available radio resources. RRM manages the assignment, re-assignment and release of radio resources, taking into account single and multi-cell aspects. The RRM, at RRC is performed by the following functions: Radio Admission Control (RAC) and Radio Bearer Control (RBC). RAC admits or rejects the establishment requests for new radio bearers in the cell, whereas RBC builds, maintains and releases the radio bearer by taking into consideration all the radio resources of the cell. RBC maintains radio bearers of existing users and releases the bearers after each user ends its communication. A new bearer will be built only if the radio resource in the cell is surplus [34].

One of the main principles of network virtualization as a concept, is the division of the control and data planes of a system. The innovation inserted by RENEV, is based on the fact that the baseband part of the RAN nodes could be shared among different HeNBs and a common RRC layer for a specific group of them could be created; the target is to concentrate and orchestrate the control plane functionalities to serve a specific group of users. It creates a common control plane among a group of HeNBs where the available radio resources could be dynamically transferred in the network, according to the users’ demands in a holistic way. The control plane of LTE-A in the RAN nodes is concentrated in RRC protocol, which is terminated in the BS on the network side. Its main functionalities are the establishment of the connections with the users, configuration of the radio bearers and their corresponding attributes and control of mobility.

### 2.1.3.3 Resources Negotiation for Network Virtualization Algorithm

Physical resource migration among HeNBs, is necessary for covering the traffic demands of the existing users. Since a HeNB has some spare resources, it is available in order to participate to the resources negotiation process. This process is decentralized since all the existing HeNBs of a topology can participate, as soon as they have spare resources.

So in this work, we propose an algorithm called Resource Negotiation for Network Virtualization (RENEV) for resources negotiation between small cells, by the cooperation of RAC and RBC functions, belonging in RRC of different small cells. This solution is based on...
radio physical resource transfer in isolation and on-demand basis. Furthermore, our solution supports common RRC scheduling between different HeNBs. The following steps describe RENEV:

- Step 1: If the resources of the serving HeNB can serve a user then it gets served [4].
- Step 2: Otherwise:
  - The user enables the RRC connection with the serving HeNB.
  - This HeNB finds the nearest\(^1\) and less loaded neighbour HeNB.
  - When it finds it, the two RRC functions of the node are enabled; the RAC function is responsible for checking if the node has the available resources and the RBC for establishing the radio bearer.
  - A control connection is created between the involved HeNBs via the logical point-to-point X2 interface.
  - The serving HeNB leases the demanded resources so the user is getting served. This happens by setting up X2 interfaces and resetting the link resolving security issues for the exchange of HeNB configuration data over the link.

The target metric that this algorithm improves is the aggregate system throughput, since the resources negotiation in terms of RBs affects the data rates that are delivered to all terminals in a system.

### 2.1.4 Analytical Formulation

#### 2.1.4.1 A. System Model and Assumptions

The general topology where the RENEV is applied is a dense residential area including two tiers; a macro BS and randomly located small cells. The two tiers use different carriers so as not to have interference issues. The algorithm is applied only in the small cell tier and consequently only users connected to that tier are considered. The users of the system have specific characteristics in terms of demanded traffic. To represent the network random deployment, it is assumed that HeNBs are randomly distributed on a two-dimensional Euclidean plane \( \mathbb{R}^2 \). Within this model, let us denote by \( x \in \mathbb{R}^2 \) the location of an HeNB.

\(^1\) The term "nearest" indicates the neighbour HeNB located geographically closer to the serving HeNB. This fact restricts the effect of the algorithm geographically, in order to avoid instability issues.
By the above considerations and notation, the signal strength from an HeNB received at location \( y \in \mathbb{R}^2 \), expressed in dB, is thus:

\[
p_i(y) = P_{Tx} - L_i - \{X_i\}, \tag{1}
\]

where \( P_{Tx} \) is the constant that includes antenna gains and transmitted power, and \( L_i \) is the path loss from the HeNB to \( y \in \mathbb{R}^2 \). Although \( L_i \) depends on the location \( y \), it is omitted for the sake of simplicity. In a general sense, the fading includes medium-scale variations which are due to shadowing, and short-scale fluctuations whose main effect is this on the bit level performance. In our case, only medium-scale variations are considered [5]. Slow shadowing is commonly described as a log-normal distributed random variable. Accordingly, the shadowing may be modelled by a normal variable, namely \( \{X_i\} \), when expressed in logarithmic units, with 0 mean and a standard deviation \( \sigma_x \) typically around 8dB.

The signal quality of HeNB \( i \) expressed in terms of signal-to-interference-plus-noise ratio (SINR) received at \( y \), when no interference is received, is given by:

\[
Q_i(y)_{dB} = p_i(y) - N_0, \tag{2}
\]

where \( N_0 \) represents the noise average power. In our work, as subcarriers are not reused by neighbouring cells, we consider that interference is negligible. In the following, the aggregate system throughput is analysed, with respect to a nominated user located at the origin. Note that in case of no ambiguity, we will omit the location variable \( y \) in involving definitions for notational simplicity.

2.1.4.2 Problem Setup

In principle, it should be beneficial to the network throughput to associate each mobile to the HeNB from which to obtain the best signal quality. In our case, we simply assume that the number of RBs per HeNB are allocated a priori and that each mobile is always connected to the HeNB of the best signal quality. This provides the upper bound of the network throughput. Let \( Y \) be the best SINR received from the HeNBs located in an area \( B \). Following (2), it is expressed:

\[
Y = \max_{x_i \in B} Q_i
\]

Discrete adaptive Modulation and Coding Schemes (MCSs), (i.e., quaternary phase-shift keying (QPSK), 16 quadratic-amplitude modulation (QAM), and 64-QAM) are supported in this work [37].
2.1.4.3 Formulas of Aggregate Throughput

Using the above definitions, we are able to formulate and derive the aggregate throughput for both downlink in the cases whether the algorithm is applied or not. In this section, the final corresponding elaborated equations are represented. Wherein the next analysis follows. In a scenario where the proposed algorithm is not applied, the users connected to a particular BS must be served with only the resources allocated to the HeNB. On the contrary, when the proposed algorithm is applied, resources can be transferred from one HeNB to another, according to their subscribers’ needs. Given a total number of users in the scenario, the throughput achieved in HeNB when the proposed algorithm is not applied, $T_i$, may be expressed as:

$$T_i = \min(N \cdot P(Q_i = Y) \cdot R, T_i^{RB} \cdot B_i),$$  

where $R$ is the mean user demand (in bps), $P(Q_i = Y)$ defines the probability of having the maximum value of SINR, $T_i^{RB}$ is the average throughput per RB in HeNB, and $B_i$ is the resource allocated to HeNB $i$ (i.e., the number of RBs). The aggregate throughput of the scenario is then,

$$T = \sum_i T_i.$$  

On the contrary, when the resources negotiation algorithm is applied, the whole pool of resources is dynamically distributed among the HeNBs. In this context, the mean throughput of a single RB within the scenario, $T^{RB}$, is denoted as:

$$T^{RB} = \sum_i T_i^{RB} \cdot P(Q_i = Y),$$  

and accordingly the aggregate system throughout is:

$$T = \min(N \cdot R, T^{RB} \cdot \sum_i B_i).$$

2.1.4.4 Analysis

According to each user’s distance from the serving HeNB centre, it obtains a certain SINR value. This value is responsible for the corresponding MCS that will be used among the available ones (i.e., QPSK1/2, QPSK3/4, 16-QAM1/2, 16-QAM3/4, 64-QAM2/3, 64-
QAM3/4). The probability of using a certain MCS in the HeNB $i$ is expressed as $P_i(MCS_n)$. The average throughput per RB in the HeNB can be calculated by the following formula:

$$T_i^{RB} = \sum_{n=0}^{5} RB^n P_i(MCS_n), \quad (7)$$

where $RB^n$ denotes the throughput achieved with a single RB according to the corresponding $MCS_n$. The terms $Q_{min}^{n}$ and $Q_{max}^{n}$ denote the minimum and maximum SINR with which $MCS_n$ is applied. Since the probability of achieving a certain MCS depends on the SINR, it could be expressed as:

$$P_i(MSC_n) = \frac{P(Q_{min}^{n} \leq Q_i < Q_{max}^{n} | Q_i = Y)}{P(Q_i = Y)}, \quad (8)$$

Where

$$Q_i = P_{Tx} - L - \{X_i\} - N_0$$

according to (2).

The denominator of the expression (8), equal to:

$$P(Q_i = Y) = P(\cap_{j \neq i} Q_i > Q_j) = \prod_{j \neq i} P(Q_i > Q_j), \quad (9)$$

since SINR values of a particular HeNB are considered to be independent from the SINR of the other HeNBs. From (9) it is necessary to calculate the probabilities of having higher SINR values in an HeNB, denoted by $P(Q_i > Q_j)$:

$$P(Q_i > Q_j) = P(P_{Tx} - L_i - \{X_i\} - N_0 > P_{Tx} - L_j - \{X_j\} - N_0)$$

$$= P(\{X_j\} > \{X_i\} + \mu_{ij}), \text{ where } \mu_{ij} = L_i - L_j. \quad (10)$$

It should be noted that $\{X_i\}$ and $\{X_j\}$ are both random variables following Gaussian distributions with the same mean and deviation. We denote as $f_X$ and $F_X$ the corresponding Probability Density Function (PDF) and Cumulative Density Function (CDF) of the shadowing expressed in dB. Based on the analysis in [6], and after a convenient change of variables, equation (10) is equal to

$$F_{X_i}(\frac{\mu_{ij}}{\sigma\sqrt{2}})$$
Furthermore, the probability of the joint occurrence of pairwise and global independent events is equal to the product of the events’ marginal probabilities. Correspondingly, the numerator of (8) can be calculated as follows:

\[ P(Q_{n}^{\text{min}} \leq Q_i < Q_{n}^{\text{max}} \cap Q_j = Y) = \prod_{j \neq i} P(Q_{n}^{\text{min}} \leq Q_i < Q_{n}^{\text{max}} \cap Q_j > Q_j). \]  

If we substitute the values \( \{X^0_i\} = P_{TX} - Q_{n}^{\text{max}} - L_i \) and \( \{X^1_i\} = P_{TX} - Q_{n}^{\text{min}} - L_i \), each of the components of (11) can be expressed as follows:

\[
P(\{X^0_i\} \leq \{X_i\} < \{X^1_i\} \cap \{X_j\} > \{X_i\} + \mu_{ij}) = \int_{\{X_i\} + \mu_{ij}}^{\infty} \int_{\{X_i\}}^{\{X_j\}} f_{X_i}(x_i) f_{X_j}(x_j) \, dx_j \, dx_i
\]

\[
= \int_{\{X^0_i\}}^{\{X^1_i\}} f_{X_i}(x_i) \, dx_i - \int_{\{X^0_i\}}^{\{X^1_i\}} F_{X_i}(x_i + \mu_{ij}) f_{X_i}(x_i) \, dx_i
\]

\[
= (F_{X_i}(\{X^1_i\}) - F_{X_i}(\{X^0_i\})) - \int_{\{X^0_i\}}^{\{X^1_i\}} F_{X_i}(x_i + \mu_{ij}) f_{X_i}(x_i) \, dx_i
\]

\[
= (F_{X_i}(\{X^1_i\}) - F_{X_i}(\{X^0_i\})) - V(\{X^1_i\}, \{X^0_i\}).
\]  

(12)

Also, as noted, \( f_{X_i} \) and \( F_{X_i} \) are the PDF and the CDF of a normal variable. Expression (12) depends on the values of the random variables \( \{X^1_i\} \) and \( \{X^0_i\} \), as well as it has no closed form and is evaluated numerically.

### 2.1.5 Performance Evaluation

We consider 3GPP HeNB settings for the setup of small cell networks. The system transmission bandwidth is equal to 20 MHz, corresponding to 100 RBs, and the transmission model is Single Input Single Output (SISO). A custom-made Matlab simulation tool is employed to validate the proposed scheme. Consider that there are 6 small cells not uniformly distributed in an area of 100 m x 100 m, belonging to one service provider and one network operator. The propagation path loss in the small cell network is given by the following path loss model [7]:

\[
L_{dB} = 37 + 30 \log_{10}(d) + 18.3 f^{\frac{f + 2}{f + 1} - 0.46},
\]  

(13)
where $d$ is the distance in meters from the antenna and $f$ is the number of penetrated floors in the propagation path. For dense wireless networks that could be located generally, including outdoor urban areas where there are less penetrated walls and floors, $f=3$ is considered in our work. It is assumed that the total transmit power including the antenna gain of each HeNB is $32$ dBm. Shadow fading is modeled as a random variable with log-normal distribution of mean and standard deviation $8$ dB. The received noise power is the one of an Additive White Gaussian Noise (AWGN) channel. In the following, we investigate small cell network performance under random topology.

In this set of experiments, we compare the system with and without the application of RENEV, to illustrate the benefits gained in terms of network's aggregate throughput. We also compare it with NVS, another framework presented in works [24] and [25] of the state of the art, that opportunistically allocates the unused resources among the existing slices in a BS. We adapt this framework to our scenario that consists only of HeNBs, by creating distinct slices, each one accommodating a certain percentage of the overall RBs that can serve a specific number of users. All users inserted into the topology, download files using File Transfer Protocol (FTP) at an average data rate of $300$ Kbps in the downlink. Following common practice in commercial cellular networks, FTP requests are always admitted regardless of the system’s load conditions.

![Figure 1. Aggregate System Throughput for different number of offered loads.](image)

In Figure 1 the aggregate system throughput is shown with respect to an increasing offered traffic load for the system with and without the application of RENEV as well as for the NVS.
framework. For low offered load, up to 9Mbps, the system’s behaviour is the same; the users’ demanded traffic is served in all the cases. However, as the load increases, the system without the application of RENEV is able to serve less traffic load, compared to the system where the algorithm is applied. When saturation is reached (i.e. when the offered load equals 27Mbps) the achieved throughput raises 45.11%. This could be explained by the fact that in the first case, the available resources are distributed among the group of the participants HeNBs in order to cover the maximum of the users’ traffic demand.

In the case where RENEV is not applied, each HeNB controls its own resources and after a while these resources are depleted. System saturation is reached in the case where more load is introduced but the existing RBs are depleted and consequently no more users can be served. NVS reaches higher system throughput than the system without the application of RENEV, due to its capability to allocate the free resources in the slices that contain users that need it. Since here one type of traffic and fixed percentage of resources among the slices are introduced, this solution restricts the number of the resources that are transferred. We could consider that RENEV adds one more dimension to the vision of NVS in order to achieve virtualization in LTE-A environments. NVS is based on the heterogeneity of services and RENEV on the idea of resources transferring in HeNBs that do not own specific percentage of resources to transfer. So, RENEV takes advantage of this capability adding one more degree of flexibility in the resource transferring among the existing flows that share one or more physical BSs.

Figure 2 depicts the percentage of the transferred RBs versus the offered load during the application of RENEV. When the demanded traffic reaches 27Mbps the system requires the highest number of RBs in order to satisfy the existing users; 23.15% of the total RBs belonging to the system are transferred. After this point, although the number of users that require resources is augmented, the number of transferred resources decreases because the system runs out of resources since all of them are already allocated to the existing users. HeNBs are only capable of transferring resources to other HeNBs when they have unused resources. Accordingly, when the offered traffic grows, the possibility of transferring resources to other cells falls. In employing RENEV, as the offered load increases, the HeNBs request more RBs. However, after a certain point, the successful RB transfer decreases.
Figure 2. Percentage of transferred Resource Blocks.

RENEV is a decentralized proposal for transferring resources among several HeNBs. This means that when the offered load is augmented, the total number of resources is distributed among the users according to their requirements dynamically as they come from a common pool. This leads to a peer-to-peer common RRC scheduling between the participants HeNBs and also a common control plane for the RAN nodes. With the use of RENEV, all the system’s resources are dynamically used according to the users' needs on an isolated and on-demand basis. In this way, the majority of the users is served, as long as spare resources exist. In any other case, the users would receive lower quality of service or they could not get even served at all.

2.1.6 Conclusions

In this section, we propose RENEV, an algorithm for resources negotiation, in LTE-A cellular environments consisting of small cells. The proposed scheme takes advantage of the fact that in such environments not all the subcarriers are used simultaneously among the cells, so spectrum resources can be transferred from one HeNB to another. Due to the fact that the resources negotiation is conducted in an on-demand and isolated way, RENEV could be regarded as a RAN virtualization solution dedicated to LTE-A systems. The proposed solution achieves significant gains in terms of system aggregate throughput.

In addition, we note potential extensions of this work. A future step is to increase the small cell density and in parallel extend RENEV so as the macro BS to participate. Something that
we should consider in this case is the potential interference between the two tiers. Besides, it would be interesting to see how the decision of which RB should be used has an impact on the achievable throughput. Finally, one interesting extension of this work is to investigate the system’s behaviour for emerging data traffic patterns which are different from the ones presented in this section, for various numbers of service and network providers.

2.2 Scalable RAN Virtualization in Multi-Tenant LTE-A HetNets

2.2.1 Introduction

Research conducted in the last years reveals that cellular networks will have to become heterogeneous and denser to meet such envisaged demands [8], thereby paving the way for the future Ultra-dense networks [9]. Considering that operating infrastructure is a significant cost for operators, the densification of the access networks and the necessity to reduce the costs will undoubtedly lead to cooperation between them and, consequently, to the sharing of resources, including infrastructure sharing itself. In this context, the provision of solutions enabling the creation of logically isolated network partitions over shared physical network infrastructure should allow multiple heterogeneous virtual networks to coexist simultaneously and support aggregation of resources. This concept defines the principle of network virtualization [9] and explains why RAN virtualization has emerged as one of the key aspects of the future cellular LTE-A networks.

Today’s cellular networks have relatively limited support for virtualization. Thus, although the 3GPP standardizes necessary functionalities to enable several core operators to share one RAN [11], neither a detailed implementation of radio resource allocation nor the mechanisms to exploit the network heterogeneity of the dense multi-tier architectures, defined in the latest release of LTE-A, are provided [12][13]. Moreover, the dynamic composition of these coexisting and isolated heterogeneous virtual networks is not defined. Therefore, the particular definition of algorithms that implement RAN virtualization still remains an open issue.

The main challenges that should be addressed by RAN virtualization in the framework of LTE-A are i) the capacity limitation imposed by resource allocation, ii) the complete isolation between multiple coexisting services, and iii) the additional signaling overhead of each proposed solution. These challenges are even more complex to tackle in dense multi-tier scenarios, where Small Cells (SCs)[12], are characterized by reduced coverage areas and therefore make the scenario more prone to geographical traffic non-uniformities [14]. Traffic
load and deployment are the foremost aspects of investigating the potential effectiveness of RAN virtualization. Although research solutions proposed so far have been mainly focused on the virtualization of resources in each BS\(^2\), there is still a gap in the literature for solutions that can abstract the available resources to deliver them to multiple tenants, considering the geographical traffic variations that can occur in heterogeneous scenarios.

This work, taking into account the gaps in the current literature, is aimed to shed light on the limitations of the existing LTE-A RAN virtualization solutions by coping with dense multi-tier networks. Specifically, our contribution is twofold. Firstly, we extend and modify our previous proposal from [15] (as described in Section 2.1), the Resources nEgotiation for NEtwork Virtualization (RENEV) algorithm, for dynamic virtualization of radio resources spread in a two tier topology. Motivated by the geographical traffic variations, we propose a solution where baseband modules of distributed BSs, interconnected via the logical point-to-point X2 interface, cooperate to reallocate radio resources on a traffic need basis. Our proposal is based on the concept of physical resources transfer, defined as the possibility of reconfiguring the Orthogonal Frequency Division Multiple Access (OFDMA)-based medium access of two BSs, to allow a BS to use a set of subcarriers initially allocated to another BS. Resource customization to various tenants, i.e., BSs, is conducted after appropriate signaling exchange. Secondly, we identify the basic limitations and the signaling overhead caused by the algorithm to the current 3GPP LTE-A architecture. In that sense, RENEV is harmonized and adapted to be compatible with LTE-A standard. Additionally, an insight on the analysis of the additional signaling overhead of the scheme is given, since it is a key issue for virtualization, particularly as the network planning becomes denser.

### 2.2.2 State Of The Art and Contribution

Existing literature on network virtualization can be grouped into solutions for the Evolved Packet Core (EPC) Network and the RAN [16]. This paper is focused on the RAN of an heterogeneous LTE-A deployment, which, in turn, can be also divided into two main categories: dynamic resources’ slicing and spectrum sharing.

With regard to the dynamic resources’ slicing, interesting proposals are presented in [17][18][19]. CellSlice framework is proposed in [17] to achieve active RAN sharing by

---

\(^2\) Throughout the rest of this manuscript, the term BS is used to describe either a macro eNB or a small cell (SC). The exact name of the BS is defined in all the particular cases that require the exact distinction among them.
remotely controlling scheduling decisions without modifying BS’s schedulers. Instead, in [18][19][20], the authors present software defined cellular network architectures, allowing controller applications to express real-time, fine-grained policies based on subscribers’ attributes.

As for spectrum sharing [21][22][23][24] the proposals are designed to adapt the radio interface of the eNB to traffic load variations of distinct virtual networks. This objective is achieved by allowing multiple virtual networks to share the spectrum allocated to a particular physical eNB. In [21], a hypervisor makes use of a priori knowledge to schedule Resource Blocks (RBs) to different operators. In addition, the authors of [22] evaluate several sharing options, ranging from simple approaches feasible in traditional infrastructure to complex methods requiring a specialized one. A feature called NVS, has been introduced in [23], [24] and [25]. In these works, a slice scheduler on top of the Medium Access Control (MAC) scheduler is introduced such that each operator achieves its reserved resources while enabling customized flow scheduling within its slice. Additionally, [26] extends NVS solution by investigating the provision of active LTE RAN sharing with Partial Resource Reservation (PRR). In this scheme, each slice is guaranteed a specific minimum share of radio resources to be available to the operator that owns them. The remaining common part is shared among traffic flows belonging to different operators.

Based on the presented state of the art, virtualization solutions proposed so far have been mainly focused on sharing resources within a specific BS ([17][23]). In particular, whereas in some proposals resources are dynamically sliced between services with different QoS characteristics ([21][22][24]), in other proposals the same resources are virtualized and distributed among different operators with shared access to the same BS ([25][26]). Such proposals are effective virtualization solutions to address the traffic dynamics in two aspects: service and operator dimension. In the first case, the variety of services poses challenges to resource allocation, whereas the second dimension is really interesting since the distribution of traffic between different operators is not necessarily uniform. However, none of the aforementioned proposals is able to cope with dynamics in a third aspect of traffic: the geographical dimension. HetNets are characterized by dense deployment of BSs with different transmission power and overlapped coverage areas. In these scenarios, the emergence of low-power BSs (i.e., SCs), with the objective of densifying the network, has clear impacts on the traffic load: i) the distribution of the traffic between BSs is not uniform [14][27], and ii) the variability of traffic in the short-term, particularly in SCs, is increased. As a
consequence the overall capacity of the system is usually compromised by spatial non-uniformities. Therefore, even appropriate deployments, which are static in nature, are unable to tackle optimally geographical variations of traffic.

Accordingly, an efficient use of the available radio resources can be achieved if a proper coordination / negotiation of resources is carried out among the BSs. In [15], we introduced a first approach of RENEV and applied it in a deployment consisting only of SCs (i.e., HeNBs). In such environments, RENEV is responsible for reallocating / transferring radio resources by reconfiguring the OFDMA based radio interface in a decentralized manner. The innovation in [18] lies in the fact that the baseband part of the BSs is shared and a common Radio Resource Control (RRC) layer for a specific group of BSs is created in a coordinated way. RENEV is essentially designed to reconfigure the radio resources of two BSs in order to adapt the allocation of resources to the dynamics of the traffic. Thus, when there is a tenant BS without enough resources to serve the offered traffic, RENEV should find out if there are unused resources in other neighboring BSs, check if the unused resources could be reallocated, and finally reconfigure the medium access of the two BSs to reallocate them from one to the other (hereinafter also known as transfer of resources). The key question that arises concerns the role of nodes of each tier; should they work in a hierarchical manner or not. This is crucial in a resources transfer procedure, since the ability to transfer resources depends on the characteristics of each node, thereby adding more degree of complexity. RENEV offers a complementary solution to the state of the art and covers gaps found therein by introducing a new dimension in RAN sharing. Accordingly, we extend the proposal in [8], by modifying it, allowing BSs that belong to two tiers to reallocate underutilized spectrum to other BSs in order to manage geographical variations of traffic conditions. Our main contributions with respect to the state of the art can be summarized as follows:

- We introduce RENEV as a solution, that can be employed on former spectrum sharing proposals (e.g., NVS [25] and PRR [25]), in HetNet scenarios composed of two tiers, each one operating on different sets of subcarriers, when geographical traffic non-uniformities render the initial allocation of resources insufficient. RENEV reallocates the resources among tenant BSs on-demand and in a decentralized manner, while the virtualization solutions proposed so far in the literature (e.g., NVS [25] and PRR [25]) only allow resources sharing among distinct services and/or operators in a BS basis.
We demonstrate that RENEV could be applied on top of existing virtualization solutions (e.g., NVS [25] and PRR [25]), thereby guaranteeing its operation in multi-service multi-operator scenarios [22]. The implementation of RENEV does not impose additional con- strains to the virtualization of resources within each particular tenant BS proposed by the aforementioned solutions.

We analytically derive the upper bounds of the throughput with and without RENEV.

We provide the detailed description and the analytical model of the signaling introduced by RENEV. This analysis arises as a key point in the dimension of the physical connections that support the logical X2 interface.

2.2.3 RENEV in a Heterogeneous Scenario

2.2.3.1 Scenario under consideration

In this subsection, we introduce the characteristics of the architectural elements of the scenario, i.e., the RAN nodes (BSs) and the control nodes as well as their interconnecting interfaces. Regarding the RAN elements, the underlying considered network is a residential region composed of a number of open access mode SCs placed throughout its coverage area in clusters, close to each other, in random positions [28] (see Figure 3).

![Figure 3. Heterogeneous two-tier deployment with SCs deployed in clusters.](image)

The two tiers are initially assigned disjoint frequency bands [29]; however, by exploiting the concept of Carrier Aggregation (CA), both tiers can operate on the whole bandwidth [12]. Most RAN nodes maintain standardized connections to each other, for example, BSs are connected to their neighbors using the point-to-point, logical X2 interface to support direct control and data information exchange. This interface is designed to logically connect two BSs (i.e., two eNBs, one eNB and one SC, or two SCs) [13]. Furthermore, we focus on the downlink, where the RB is the basic time-
frequency resources unit. In principle, any RB can be assigned to one or several BSs subject to interference constraints. The eNB is assumed to transmit with a fixed power per RB. The downlink transmitted power per RB is also fixed and equal among the SCs [14].

The BSs are connected to the EPC directly with the Mobility Management Entity (MME) or through an intermediate node, named Home eNB Gateway (HeNB GW) using the S1 interface [31]. These entities constitute the control nodes that manage BSs to provide a radio network. Based on [13], three ways of interconnection of the tenant BSs arise: (i) a cluster of SCs (i.e., in our test case HeNBs) connected to the same HeNB GW, (ii) a group of eNBs connected to the same MME and (iii) a group of eNBs as well as SCs associated to the same MME. In the first and second case, the HeNB GW and the MME concentrate the control plane of the SCs and the eNBs respectively. The last case is considered the most general one, since the MME integrates the control plane of both types of BSs within a certain geographical area. Despite the different cases presented in [13], from a BS’s perspective all cases are identical in terms of signaling. Therefore in a scenario like this, we further assume that the involved BSs are necessarily deployed over the same geographical area, and therefore connected to the same control node (i.e., MME).

With regard to the owner operator of the aforementioned architectural elements, it should be clarified that there can be more than one owning control nodes in a geographical area. The network operator owning the control node is defined by the existing Service Level Agreement (SLA). Therefore, there can be operators only owning BSs or control nodes, and operators owning both kind of nodes in the same geographical area. This fact implies that the scenario under study is valid both for a single as well as for multiple operators.

**2.2.3.2 Radio Resource Management Functions**

The management of spectrum resources allocated to the BSs of the scenario under consideration relies on their control plane. The control plane of a BS in LTE-A is logically divided in two entities: baseband and network module, as defined in the standard in [25]. The former is responsible for bearer setup, to register users to the network via RRC protocol, whereas the latter connects the BS with the EPC. Radio Resource Management (RRM) is implemented in baseband module of a BS with primary goal to control the use of radio resources in the system, by ensuring QoS requirements of the individual radio bearers and minimization of the overall use of resources.
Focusing on the baseband module, two fundamental functions of the RRM jointly manage the resources of a BS: the RBC and RAC[13]. On the one hand, RBC is responsible for the establishment, maintenance and release of radio bearers. When setting up a radio bearer, RBC considers the overall resource situation and QoS requirements of in-progress sessions [32]. Correspondingly, it is involved in the release of radio resources at session termination. On the other hand, the task of RAC is to admit or reject the establishment requests for new radio bearers. RAC ensures high radio resource utilization by accepting bearer requests as long as radio resources are available. At the same time, it ensures proper QoS for in-progress sessions by rejecting radio bearer requests when they cannot be accommodated [33]. A new bearer will be built only if radio resource in the cell is available and it will be released after the end of users’ communication [34]. Based on the role played by RBC and RAC functions in the management of the radio resources, any RRM technique aimed to improve the efficiency in the dynamic allocation of the radio resources among BSs must interact with these two functions.

2.2.3.3 Proposed Algorithm: RENEV for HetNet LTE-A deployments

In dense heterogeneous scenarios, as described in Section 2.2.3.1, traffic non-uniformities among BSs make resource allocation a challenging task. A dynamic coordination of radio resources is required to address such kind of variations. This is the objective of RENEV in these environments; reallocating resources in terms of RBs, to satisfy new incoming user requests in tenant BSs.

Let us define the number of RBs initially allocated to a particular BS as RB, and the number of RBs required to serve the demand of its associated users as u. By definition, the number of available RBs in this specific BS, denoted as r, can be expressed as \( r = RB - u \). As long as \( r > 0 \), the tenant BS will be able to serve the offered traffic. Conversely, when \( r < 0 \), the BS will start to degrade users’ performance and block UEs’ incoming attachment requests.

It is particularly worth noting that in HetNets the significant variability of the traffic among neighbouring BSs can lead to the paradox of having BSs with \( r < 0 \) and, at the same time, BSs with \( r > 0 \). RENEV is defined as the decentralized procedure intended to match the tenant BSs with \( r < 0 \) and the ones with \( r > 0 \), and manage the exchange of control messages to reconfigure the allocation of resources among them. For this reason, RENEV is divided into two sequential phases, as shown in Figure 4.
First, the detection phase, where a BS with $r < 0$ seeks among the neighbouring BSs if any of them has $r > 0$. This search is carried out by polling one by one the neighbouring tenant BSs to figure out the amount of available resources (i.e., $r$). Subsequently, the second phase, which is referred to as transfer phase, is only executed if the tenant BS with $r < 0$ finds neighbouring BSs with $r > 0$. This phase consists in re-configuring the two involved BSs.

The details of each phase are stated below, and a proposal of the messages exchanged during the two phases is described in Section IV. However, and before proceeding with the details, it is worth describing the basic nomenclature used hereafter:
• Serving BS: is the node that a User Equipment (UE) is associated to and it is responsible for serving it.

• Requesting BS: is the node that, after receiving an access request from a UE, determines that the request cannot be accommodated with the available resources. It is precisely at this time, that the node takes the role of Requesting BS and triggers a requesting process among the neighboring BSs to figure out if there are unused resources.

• Requested BS: is the node that, after a neighboring Requesting BS triggers a requesting process, receives a request to inform about its unused resources.

• Donor BS: is the node that, upon the completion of a requesting process triggered by a neighboring Requesting BS, is selected to transfer resources to the aforementioned Requesting BS.

• Recipient BS: is the role taken by a Requesting BS after reconfiguring the radio interface to use the resources transferred from a Donor BS.

Since, in general, spectral efficiency of SCs is higher than spectral efficiency of eNBs, SCs play the role of Requesting BSs. On the other hand, the role of Requested BS can be held either by SCs or an eNB. In that sense, SCs can be both Donor and Recipient BSs, whereas eNB is always a Donor BS.

I. Detection phase: If a user can be served by resources owned by the Serving BS (i.e., \( r > 0 \)), then it is served [26]. Otherwise, the Serving BS, after setting up a RRC connection on the air interface with the user requiring service provision, triggers RENEV by adopting the role of Requesting BS. At this point the detection phase starts (see the light shaded area in Figure 4). Next, the Requesting BS scans the local network\(^3\) to find a potential Donor BS by polling BSs within it. The polling procedure undertaken by the Requesting BS may itself be divided into two steps. First, the Requesting BS polls each neighbouring SC, one by one, to monitor the resources status of the SCs tier. Secondly, if and only if there are not available resources in the SCs tier, the Requesting BS polls the macro eNB. After completing the requesting process, the Donor BS is selected among the set of Requested BSs according to the following criteria:

---

\(^3\)The local network of a BS is defined as the set of BSs deployed in its vicinity. Generally, this local network consists of an eNB and a finite number of SCs under its coverage area, as shown in Figure 3.
1. **Load:** The Requested BS with more unused resources is selected as the Donor BS. Yet, in order for the Donor BS to be able to accommodate possible further increase of the traffic demand in the short/mid-term future, a Requested BS can only become a Donor BS if the amount of remaining resources after the transfer is above a minimum threshold.

2. **Proximity:** For a set of Requested BSs likely to become the Donor BS, and if more than a single Requested BS has the same amount of unused resources, the Donor BS will be the BS with the minimum distance to the Requesting/Recipient BS. This criterion guarantees that the effect of the algorithm is geographically restricted to limit undesirable instability problems caused by the nature of the wireless medium.

Regarding the implementation details of this phase, when a user is attached to the Requesting BS, the RRC connection establishment is used to make the transition from RRC Idle to RRC Connected mode. This transition is carried out before transferring any application data, or completing any signaling procedures, as shown in Figure 4. RRC establishment procedure is always initiated by the user but it can be triggered by either the user or the network ([26], [27]). Also, when the Requesting BS scans the network to find a Donor BS, a coordinated control connection of their baseband parts is created via the logical point-to-point X2 interface. Every time that a polling procedure between a Requesting and a Requested BS is carried out, two messages are exchanged through X2 interface (one from the Requesting BS to the Requested BS, and another one vice-versa).

**II. Transfer phase:** Upon detecting the Donor BS, the transfer of resources from the Donor BS to the Recipient BS takes place via X2 interface. It is worth noting, that the exchange of BS configuration data over the link must be preceded by resetting the link resolving security issues. In the proposed scheme, RAC and RBC functions, belonging to RRC layer of distinct neighbouring BSs, cooperate to provide seamless service to the end users (first action of the transfer phase, dark shaded in Fig. 2). We leverage the logical split of a BS into baseband and network modules and create a common RRC process among the Recipient BS and the Donor BSs. When the Requesting BS finds the Donor BS, RRC functions of the two nodes are enabled; RAC is responsible for checking if the node has available resources and RBC for establishing the radio bearer; it is in that moment that the Requesting BS becomes
the Recipient BS. Finally, the Donor BS leases the demanded resources, which are used by the Recipient BS. During this phase, the medium access of two involved BSs is reconfigured and spectrum is lent by the Donor BSs through the control communication of the nodes. This process is seamless to end users since RRC connection is maintained with the initial Requesting BS.

It is worth pointing out that X2 resource transfer process is carried out only between the two involved BSs, without the participation of additional BSs or gateways. The resources transfer is successful when both BSs are reconfigured. The details of all exchanged messages are explained in Section 2.2.4.

2.2.4 Signaling Design Considerations

The additional signaling overhead introduced in the network is a key aspect of the proposed solution, since it could limit or constrain its feasibility. This section is intended to analyze in detail the signaling messages exchanged in the network to implement RENEV, as well as the compliance of the algorithm with the current standards and architecture of LTE-A.

Any procedure concerning the accommodation of a new user in a cell, starts with its attachment as explicitly defined in the standard [33][34]. The attachment of a user to a new cell is characterized by two main processes: firstly, the communication between the UE and the Serving BS over air (i.e., Uu) interface, and secondly, the communication between Serving BS and the MME to exchange initial UE context setup over S1 interface.

With regard to the message exchange over Uu interface, the corresponding user sends the attach request message to the Serving BS, as also defined in the standard [26], [27]. After sending the first message of random access procedure (i.e., specific pattern/signature) to the network, denoted as RACH preamble, RRC connection is established. The initial UE context setup, consists of an exchange of messages with the purpose of transferring UE context information from the MME to the Serving BS. These messages are exchanged over S1-AP application layer using SCTP. When the appropriate RRC transport container is received by the Serving BS, the establishment of a dedicated SCTP control stream on S1-MME is triggered [35].

The described procedure is nonetheless subject to the availability of resources in the Serving BS. In that sense, RENEV aims to transfer resources from one BS to another to minimize the number of unsuccessful procedures. Therefore, RENEV should be executed after the UE attachment request and before the UE context exchange. As several nodes are
involved into the procedure, and for the sake of clarity, we have divided the call flow of the required messages that should be included to the standard in two sub-figures, depicted in Figure 5. Thus, Figure 5(a) shows RENEV and the previous UE attachment procedure (defined in the standard), whereas Figure 5(b) depicts RENEV and the subsequent UE context exchange procedure (defined in the standard).

![Figure 5. Call Flow of the messages for (a) UE Attachment and RENEV and (b) UE Context Exchange and RENEV.](image)

The direct communication between two BSs is conducted via X2, using the X2 Application Protocol (X2-AP) [13]. X2-AP messages are characterized by communication context identifiers and some specific parameters called Information Elements (IEs). These define the source and target BS, as well as characteristics of the transferred message. In the following, we introduce details of the exact messages required to implement RENEV.

### 2.2.4.1 Detectionphasesignaling

When applying RENEV, the first process to carry out includes the polling procedure to detect spare resources (see Figure 5(a), messages 1, 2 and 3). During this operation, the Requesting BS scans the network to find the Donor BS, as shown in Figure 4(in the light shaded box). For each Requesting BS-Requested BS pair the polling process entails the information exchange about resources and load status [13]. In the standard, the X2-AP defines two Elementary Procedures (EP) for this same purpose, namely the “Resource Status Initiation” and “Load Indication” procedures [13]. The former is defined as a class 1 EP (i.e., it consists of two messages, a request and a response, namely “X2-AP:RESOURCES STATUS REQUEST” and “X2-AP:RESOURCE STATUS RESPONSE“
messages), whereas the latter is defined as a class 2 EP (i.e., it consists of a single message, without response, namely “X2-AP:LOAD INFORMATION” message). RENEV makes use of these two EPs, defined by the X2-AP, to implement the detection phase.

As shown in Figure 5(a), the Requesting BS sends the standardized “X2-AP:RESOURCE STATUS REQUEST” message to the Requested BS (Figure 5, message 1) asking for the following information (known as IE in the X2-AP nomenclature): the percentage of RBs in use, the load on S1 interface, and the hardware load. The Requested BS returns a response and then reports each IE for both uplink and downlink with the standardized “X2-AP:RESOURCE STATUS RESPONSE” message (Figure 5, message 2). Further details of these messages can be found in [36]. Also, Load Indication procedure is used to transfer interference coordination information between neighboring BSs managing intra-frequency cells. The standardized “X2-AP:LOAD INFORMATION” message (Figure 5, message 3) includes three IEs for the controlling cell: the transmitted power in every downlink RB, the interference received in every uplink RB, and the list of uplink RBs in which the BS intends to schedule data stations for mobiles [36]. These control messages are necessary before transferring additional coordination information for establishing a common RRC layer among BSs with RENEV. This procedure is repeated for all neighboring SCs. If none of the requested SCs has enough unused resources, the procedure is repeated with the eNB. Up to this point, all messages used by RENEV in the detection phase are defined in the standard [36].

2.2.4.2 Transfer phase signaling

In RENEV, the transfer of resources is defined as the reconfiguration of a set of unused subcarriers to be vacated by the Donor BS and subsequently used by the Recipient BS. As this procedure is not considered in X2-AP, a new Class 1 EP compatible with the standard should be defined. In this paper, the two proposed messages of the new EP are the messages 4 and 5, as depicted in Figure 5. We denote them as “X2-AP:METASIGNALLING INFORMATION REQUEST” and “X2-AP:METASIGNALLING INFORMATION ACKNOWLEDGE”, although other possible implementations are not precluded.

Once the Donor BS is selected, the initiating “X2-AP:METASIGNALLING INFORMATION REQUEST” message (message 4 in Figure 5) is transmitted from Requesting BS to the Requested BS to show that resources are required by the former. The message must contain the following IEs: Message Type, Requesting BS X2-AP ID, Requested BS X2-AP ID and the corresponding transparent container. These IEs indicate
the number of necessary RBs to cover the needs of the corresponding UE, as well as the identities of the Requesting and Requested BSs. For its part, the Donor BS returns a response to the Recipient BS via “X2-AP:METASIGNALLING INFORMATION ACKNOWLEDGE” message (see Figure 5, message 5). This message carries all control information needed to execute the actual transfer of resources. The corresponding IEs are the Message Type, Cause, Bearers Admitted List, Bearers Rejected List and the equivalent transparent container. All these IEs are necessary to confirm that the requested RBs exist in the Donor BS and that they are available for use by the Requesting BS.

2.2.5 Throughput Analysis

2.2.5.1 System Model

As described in Section 2.1.3.1 the scenario consists of a single macro eNB (hereafter denoted as BS0 and located at the center of the scenario) and a SCs tier, made up of a set of SC clusters, each one consisting of \( N \in \mathbb{N} \) SCs (denoted as \( BS_i \), with \( 1 \leq i \leq N \)), randomly distributed on a two-dimensional Euclidean plane \( \mathbb{R}^2 \). As clusters are not overlapped, there is no loss of generality in assuming one SC cluster within the eNB coverage area, creating a set of \( N + 1 \) BSs, which is referred to as \( B = \{BS_i : 0 \leq i \leq N\} \).

We denote as \( X \in \mathbb{N} \) the number of the overall users within the deployed scenario. These \( X \) users are divided into \( N + 1 \) traffic layers, each one geographically spanned over the coverage area of a BS. The coverage area of a \( BS_i \) is defined as the region where users are served by this specific BS and all the users are assumed to be connected to the BS from which they receive the best SINR, given that there are available RBs within \( BS_i \). Given the described scenario, if the proportion of users contained within the coverage area of \( BS_i \) is denoted as \( \alpha_i \), the number of users within this coverage area may be expressed as \( X_i = a_i X \). Within each traffic layer, users are distributed uniformly. This analysis is focused on the downlink.

2.2.5.2 General Throughput Formulation

In LTE-A systems, the throughput achieved by a specific UE is tightly coupled with the BS to which it is connected, the SINR received from this BS and the available resources. In the following analysis, we first derive general expressions for the throughput per RB achieved in the scenario, and finally we derive the overall system throughput when RENEV is applied and when RENEV is not applied. In particular, the LTE-A standard defines a discrete
set of Modulation and Coding Schemes (MCSs) with the following possible configurations in
the downlink for data transmission for both SCs and the eNB: QPSK (1/8, 1/5, 1/4, 1/3, 1/2, 2/3, 3/4),
16-QAM (1/2, 2/3, 3/4) and 64-QAM (1/2, 2/3, 3/4) [37]. Based on a target bit error rate, the
MCS is selected by the BS according to the SINR received by the user. In that sense, given
that the transmission rate depends on the applied MCS, the expected transmission rate per
RB of a user connected to BS_i is

$$E[R_i] = \sum_k P(MCS_i = k) \cdot R_{ik}, \quad (1)$$

where $P(MCS_i = k)$ is the probability of using the kth MCS in BS_i, and $R_{ik}$ is the
transmission rate achieved within a single RB with the kth MCS. The detailed derivation for $P
(MCS_i = k)$ may be found in Appendix A. Note that (1) is valid for both eNB and SCs.
However, due to the overlapping of the coverage areas of the SCs and the eNB, the users
located within the coverage area of a SC could also be connected to the eNB in case the
available resources allocated to the SCs do not suffice to serve all the users. In other words,
a user of the ith traffic layer (with i \neq 0) could also get connected to BS0 due to lack of
resources in BS_i despite SINR_i > SINR_0 holds. Hence, if a user within the ith traffic layer
(with i \neq 0) is served by the eNB, the expected transmission rate per RB is given by

$$E[R^0_i] = \sum_k P(MCS^0_i = k) \cdot R_{ik}, \quad (2)$$

where $P(MCS^0_i = k)$ is the probability of using the kth MCS in the eNB (i.e., BS_0)
with a user within the ith coverage area (with i \neq 0). Based on this rationale, for a given
number of users $X_i$ within the coverage area of BS_i (i\neq 0) there is a group of users associated to BS_i namely $X^i_i$ and a group of users
associated to BS_0 denoted by $X^0_i$. Thus for a given $X_i$ the expected number of users
associated to BS_i is expressed by

$$E[X^i_i] = \min \left( \frac{R_{Bi}}{d} \cdot E[R_i] \right), \quad (3)$$

where $R_{Bi}$ is the number of RBs allocated to BS_i and $d$ is the specific demand of every single
user (in bps). Finally by definition $E[X^0_i] = X_i - E[X^i_i]$.

According to the definition stated above, the total throughput, expressed as the sum
of the throughput of each particular BS (i.e., $T = \sum_i T_i$), depends on the number of users
connected to each BS, the transmission rate per RB, as well as the amount and the
distribution of the available resources. In that sense, the virtualization solutions proposed in
the literature, e.g., NVS [25] and PRR [25], determine how resources are distributed within
each BS. For instance, if NVS is implemented, a particular BS will lack resources as soon as
one of the traffic flows consumes all the resources devoted to it [25]; if PRR is applied, all the
traffic load will be served as long as the shared part of resources belonging to particular BS,
is nonempty [25]. Thus, based on the aggregate traffic demand in a particular BS and how
the resources are distributed within it, the needs or surpluses of resources in a BS could vary.

Taking into account this impact, the expressions derived in the following subsections
assume the use of a first-come first-served policy in each BS of the scenario under
consideration, without prioritization of the traffic slices. Note that this policy is equivalent to
an extreme case of PRR, where 100% of the resources in each particular BS are shared and
delivered on-demand, without any previous reservation for specific traffic slices (hereafter
denoted as PRR 100%). This assumption (regardless of whether RENEV is applied or not),
results in the upper bound of the aggregate throughput. However it also poses some
limitations in the QoS required by different traffic slices.

2.2.5.2.1 Aggregate Throughput with RENEV

RENEV is characterized by distributing, in a decentralized manner, the available
resources among existing BSs. Although resources are negotiated in a peer-to-peer fashion
among BSs, the procedure can be stochastically modelled as a single pool of resources
dynamically allocated to the tenant BSs, when RENEV and PRR 100% are implemen-
ted. Thus, the resources initially allocated to the SCs tier are reallocated when and where
required, whereas the resources allocated to the eNB, after accommodating the traffic served
by the eNB, are also redistributed among the tenant SCs. Let us denote the throughput
served by the eNB and generated by the $X_0$ users, as $T_{R,0}$. This throughput will equal the
traffic generated by $X_0$ users associated to BS0, subject to the availability of sufficient
resources (i.e., $RB0$). Thus,

$$T_{R,0} = \min \left( X_0 \cdot d, E[R_0] \cdot RB_0 \right).$$

(4)

Note that, when RENEV is applied, the eNB tends to transfer resources to the SCs, if
necessary and feasible, rather than serve users within the coverage area of the SCs.
Therefore $X_0 = 0$ for $\forall i \neq 0$. In turn, SCs serve their users with all the resources allocated
within the SCS tier, as well as with unused resources in the eNB, RB0. The number of users located within the SCs is equal to $\sum_{i \neq 0} X_i = X (1 - \alpha_0)$. The unused resources in the eNB can be allocated simultaneously to more than one Requesting BSs that do not overlap within each SC cluster. The number of Requesting BSs that can reuse these resources equals the number of non-overlapped groups of SCs in each cluster. However, the demands of Requesting SCs are not equal. Thus, the portion of RBs that each $BS_i$ can use, will also depend on the proportion of users contained within the coverage area of $BS_i$ (i.e., $a_i$), as it is further calculated in Appendix B.

The application of RENEV may be modeled with two unified pools of resources; one composed of the RBs belonging to the SCs tier (denoted as $RB_T = \sum_{i \neq 0} RB_i$) and one consisting of the RBs from the eNB. Each Requesting BS will receive proportionally to its traffic load, resources from the SCs pool (i.e., $\frac{a_i}{1 - a_0} RB_T$) and the corresponding portion of resources belonging to the eNB pool denoted as $E[RB^e_i]$. Therefore, the aggregate throughput generated by the SCs tier, according to the proof provided in the Appendix B can be written as

$$\sum_{i \neq 0} T_{R,i} = \min \left( X \cdot (1 - a_0) \cdot d, \sum_{i \neq 0} E[R_i] \left( \frac{a_i}{1 - a_0} \cdot RB_T + E[RB^e_i] \right) \right).$$

Consequently, the expected overall system throughput with RENEV, is given by:

$$T_R = T_{R,0} + \sum_{i \neq 0} T_{R,i}.$$  \hfill (6)

### 2.2.5.2.2 Aggregate Throughput without RENEV

Alternatively, when RENEV is not applied (still considering a first-come first-served policy per BS, or in other words PRR 100%), there is not any mechanism to reallocate resources, and consequently all BSs can only serve users with their initially allocated RBs. Hence, and similarly to (4), the throughput of each SC is expressed as $T_{NR,i} = \min(X_i \cdot d, E[R_i] \cdot RB_i), \forall i \neq 0$.

With regard to the eNB throughput, it can be divided into two components: on the one hand, the throughput offered by the $X_0$ users within the coverage area of $BS_0$ (i.e., $T_{NR,0}$); on the other hand the traffic offered by users within the coverage area of the SCs that cannot be served by these BSs due to lack of resources (i.e., $T_{NR,SCs}$):
Therefore, the aggregate throughput without RENEV is given by,

\[ T_{NR,0}^0 = \min \left( X_0 \cdot d, E[R_0] \cdot RB_0 \right) \]  

(7)

\[ T_{NR,SCs}^0 = \min \left( \sum_{i \neq 0} E[X_i^0] \cdot d, E[R_i^0] \cdot \left( RB_0 - \frac{T_{NR,0}^0}{E[R_0]} \right) \right) \]  

(8)

Therefore, the aggregate throughput without RENEV is given by,

\[ T_{NR} = (T_{NR,0}^0 + T_{NR,SCs}^0) + \sum_{i \neq 0} T_{NR,i} \]  

(9)

### 2.2.6 Additional Signaling Overhead Analysis

The densification of the network via the deployment of numerous SCs poses new challenges in the infrastructure. Specifically, the need for a backhaul to interconnect BSs and forward both data traffic and signaling has emerged as one of the key points that could constrain the feasibility of these scenarios. Focusing on the actual implementation of RENEV, the whole communication among BSs relies on the existence and capacity of the logical X2 interface (as described in Section 2.2.5). Although this logical interface is standardized \[25\] the description of the backhaul physical infrastructure in order to support it, is left open. For such a reason, it is crucial from the infrastructure provider’s perspective to assess the additional overhead introduced in the network by RENEV. In the following, we theoretically derive the number of signaling messages exchanged during RENEV operation, as well as the expression for the percentage of successful resources’ transfer requests.

Given the system model presented in Section V-A and the nomenclature used in Section III-C, each BS may be characterized by the number of RBs initially allocated to it as well as the number of used/unused RBs for a particular number of users. Thus, let us define, the number of available resources for a specific \(BSi\) as \( ri = RBi - ui \), where \( RBi \) are the RBs initially allocated to \( BSi \) and \( ui \) is the number of RBs required to serve the demand of the users associated to \( BSi \). The number of required resources, \( ui \), will be upper and lower bounded as a function of the number of users connected to \( BSi \), their traffic demand and their received SINR. Therefore, \( ui \in [u_{i,min}, u_{i,max}] \), where \( u_{i,min} \) and \( u_{i,max} \) are the numbers of RBs being required when all the UE associated to \( BSi \) use 64QAM 4/5 (i.e., the maximum throughput per RB) and QPSK 1/8 (i.e., the minimum throughput per RB) respectively. Based on these definitions, the upper and lower bounds of available resources...
for the set of BSs of the described system can be defined as \( r_{\text{min}} = \min (R_{Bi} - u_i, \max) \) and \( r_{\text{max}} = \max (R_{Bi} - u_i, \min) \).

In this context, we define the state \( S_j = \{s_{(j,1)}, s_{(j,2)}, \ldots, s_{(j,r_{\text{max}}-r_{\text{min}}+1)}\} \), where \( s_{j,k} \in \mathbb{N}_0 \) denotes the sum of BSs with a number of available resources equal to \( (r_{\text{min}} - 1 + k) \). Note that, in terms of RENEV performance, for a specific configuration (a particular set of BSs and a distribution of UEs), the system may be unequivocally described with the state \( S_j \) and the probability of being in this state, namely \( \pi_j \).

As in RENEV, the BSs first seek for resources in the SC tier and subsequently in the eNB, we decouple the analysis into the set two steps. Focusing first on the SC tier (without considering the resources in the eNB), the system may be defined by the set of possible initial states \( S = \{S_1, S_2, \ldots, S_W\} \) and the probability of occurrence \( \pi = \{\pi_1, \pi_2, \ldots, \pi_W\} \), where \( W \) stands for the number of possible states. By definition \( \sum_{k=1}^{r_{\text{max}}-r_{\text{min}}} k = N \). According to the definitions stated above the number of Requesting BSs in a given state \( S_j \), will be equal to the number of BSs with negative \( r_i \), also expressed as

\[
\sum_{k=1}^{r_{\text{min}}} \delta_{jk}.
\]

Therefore, the expected number of Requesting BSs may be written as

\[
\mathbb{E}[n_R] = \sum_{j=1}^{W} n_R(S_j) \cdot \pi_j. \tag{11}
\]

After the operation of RENEV in the SC tier, the available resources of the Donor BSs will have been transferred to the Requesting BSs to cover their needs. Consequently, the probability of having the system in a particular state \( S_j \) after executing RENEV will vary. If we denote by \( \pi_j \) the probability of being in the state \( S_j \) after the RENEV completion in the SC tier, it holds,

\[
\pi_j' = \sum_{n=1}^{W} \pi_n \cdot p_{nj}. \tag{12}
\]

\[\text{\textsuperscript{4}}\text{The probability of being in state } S_j, \pi_j, \text{ depends on the distribution of users among the BSs (i.e., } a, \text{ as stated in Section V) and the MCS selection probability, calculated in Appendix A.}\]
where $p_{nj}$ is the probability of transitioning from state $S_n$ to $S_j$. Note that not all transitions are feasible since the redistribution of resources among SCs imposes some restrictions. Thus, $p_{nj} = 0$ if and only if $S_j$ is not contained in the set of feasible future states of $S_n$, i.e., $S_j \in F(S_n)$. The detailed definition of $F(S_n)$, according to the conditions that should hold to satisfy that $S_j \in F(S_n)$, is introduced in Appendix C.

Hence, the transition probability is given by

$$p_{nj} = \begin{cases} 
1 & : j = n, \quad F(S_n) = \emptyset, \\
\frac{1}{|F(S_n)|} & : j \neq n, \quad S_j \in F(S_n), \\
0 & : \text{otherwise},
\end{cases} \quad (13)$$

where $|F(S_n)|$ is the cardinality of the set $F(S_n)$. Although the SC stier is the first alternative for RENEV to reallocate the existing resources, not all requests can be covered with the resources of this tier. Thus, and according to (11), the expected number of successful requests (i.e., when the needs of the Requesting BSs are covered by the unused resources of the Donor BSs) in the SCs tier may be calculated as

$$E[n_s] = \sum_{j=1}^{W} n_R(S_j) \cdot [\pi_j - \pi'_j]. \quad (14)$$

As RENEV is completed in the SC stier, all feasible redistribution of resources has been successfully conducted, and the system is found in state $S_j \in S$, with probabilities $\pi_j$. However, note that $S_j$ characterizes the scenario without taking into account the resources available in the eNB, i.e., $r_0$. Therefore, in these second step of the signaling analysis, a new set of states, namely $S''$, must be defined to include $r_0$.

It should be noted that the resources inserted into the system may be distributed in different ways. For instance, if all Requesting BSs are overlapped among them, the new resources will be transferred to the SC stier only once. Conversely, if not all Requesting BS overlap with the rest of the Requesting BSs, the $r_0$ resources will be transferred more than once. Therefore, if we define the number of non-overlapping groups of Requesting BSs as $Q = \{1, 2, \ldots, M\}$, where $M$ stands for the number of Requesting BSs (for instance, for $S_j$ we have $M = n_R(S_j)$), the resources can be transferred to the SC stier $Q$ times.
Thus, for a specific state $S_j$ containing $M$ Requesting BSs, the inclusion of the resources from the eNB can lead to $M$ possible new states. Specifically, a state $S_j$ results in $M$ new states. Specifically a state $S_j$ results in $M$ new states defined as $S'' = \{s''_{t,1}, s''_{t,2}, \ldots, s''_{t,r_{\text{max}}-r_{\text{min}}+1}\}$ with $s''_{t,k} = s_{j,k} + 1$ for $k = r_0 - r_{\text{min}} + 1$ and $Q = \{1, 2, \ldots, M\}$ and $s''_{t,k} = s_{j,k}$ otherwise. This set of new states is defined for each value of $r_0$. Therefore, after the inclusion of the resources available in the eNB, the system may be described by the set of new possible initial states $S'' = \{S_1'', S_2'', \ldots, S_L''\}$ and the probability of being initially in these states. Thus, it holds that

$$
\pi''_t = \pi'_j \cdot P(Q = q|N, M) \cdot P_{eNB}(r_0),
$$

where $P(Q = q|N, M)$ is the probability of having $q$ non-overlapping groups in a cluster with $M$.

Requesting BSs out of $N$ BSs (further calculated in Appendix D) and $P_{eNB}(r_0)$ is the probability that the eNB has $r_0$ spare RBs to be transferred to the Requesting BSs. For a given scenario, the latter is a random variable that depends on the resources allocated to the eNB, the number of users and the traffic demand of each user.

Henceforth, we use the same calculation method that we used for the SCs tier to derive the expected number of successful requests. Firstly, the expected number of Requesting BSs is calculated as in (11), using the new probabilities of occurrence $\pi''$ denoted as $E[n''_R] = \sum_{j=1}^{L} n_R(S''_j)\pi''_j$. After the application of RENEV, the available resources of the eNB will have been transferred to the Requesting BSs. The new transition probabilities from state $S''_n$ to $S''_j$ for this phase according to (12) will be equal to $\pi'_j = \sum_{n=1}^{L} \pi''_n p'_{nj}$ where $p'_{nj}$ is calculated with (13) and the set of feasible future states $F(S''_n)$ according to Appendix C.

Under the conditions stated above, it cannot be assured that all requests can be covered with the resources of the eNB tier. Thus, the expected number of successful requests in the eNB tier

$$
E[n''_R] = \sum_{j=1}^{L} n_R(S''_j) \cdot [\pi''_j - \pi''_j].
$$

Therefore the total expected number of successful requests by both tiers after the completion of RENEV is equal to $E[n_{\text{total}}] = E[n''_R] + E[n'_s]$, and the probability of successful requests is calculated as $\frac{E[n_{\text{total}}]}{E[n_R]}$. 

Security: Public
The number of signaling messages exchanged by the BSs depends on the total number of BSs (i.e., \(N+1\)), the number of Requesting BSs, and the number of successful requests. In particular, and by observing Figure 5, it can be noticed that all Requesting BSs (whose number is in average equal to \(E[n_R]\)) exchange 3 messages (messages 1, 2 and 3) with the rest of the BSs. Additionally, the Requesting BSs not being able to obtain resources from the SCs exchange the aforementioned three messages with the eNB. Finally, if any of the requests is successful, the Requesting BSs exchange 2 messages (messages 4 and 5 in Figure 5). Therefore, the expected number of signaling messages exchanged by RENE may be expressed as

\[
I = 3 \cdot (N - 1) \cdot E[n_R] + 3 \cdot E[n'_R] + 2 \cdot E[n_{total}].
\] (17)

### 2.2.7 Performance Evaluation

#### 2.2.7.1 Simulation Scenario and Parameters

The number of clusters per eNB coverage area can vary from 1 to optional 4, and the number of SCs per cluster can vary from 1 to 10 depending on the actual deployment [38]. Therefore, our simulations scenario consists of an eNB overlaid with a cluster of SCs, consisting of 6 outdoor SCs (i.e., HeNBs), operating on the same carrier frequency [31], [32]. We conducted Monte-Carlo extensive simulations (with a thousand iteration to achieve statistical validity) in a custom made simulation tool implemented in MATLAB, using random deployments of an SCs cluster placed within the eNB coverage area. In each iteration, mobile users are distributed independently and non-uniformly; i.e., 2/3 are dropped within the SC tier [38], [39]. The simulation parameters are listed in Table I; the 3GPP related parameter values are based on [29]. The overall system bandwidth consists of 2 bands of 20 MHz, operating at 2 GHz, each one assigned to each tier using CA. Packet scheduling is proportional fair both at the eNB and SCs. We conduct simulations for a full buffer traffic model [38]. Users download files using FTP at an average data rate of 300 kbps downlink.

As discussed in the previous sections, RENE is a complementary virtualization solution implementable on top of existing solutions. Hence, in the scenario under consideration both NVS and PRR are simulated with and without RENE. NVS creates distinct slices of spectrum in each particular BS.
TheseslicesaccommodateequalpercentageoftheoverallRBs,eachoneresidinginaspecifictrafficflow. PRRframework,guaranteesaminimumnumberofRBspersubframeonaverageforeachtrafficflow,whichisavailablewhenaparticularflowwantstouseit(i.e.,reservedpart).

**Table 1. Basic System Parameters used in the Simulation.**

<table>
<thead>
<tr>
<th>PARAMETERS</th>
<th>SETTINGS/ASSUMPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network layout</td>
<td>Cluster of 6 SCs randomly placed per eNB</td>
</tr>
<tr>
<td>Inter-site distance/cell radius</td>
<td>Macro cell: 500 m (ISD)</td>
</tr>
<tr>
<td></td>
<td>SC: 25 m (Cell radius)</td>
</tr>
<tr>
<td>Transmit power</td>
<td>Macro eNB: 46 dBm, SC: 17 dBm</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>20 MHz at 2 GHz for each tier</td>
</tr>
<tr>
<td>Antenna configuration</td>
<td>2 x 2 MIMO with rank adaptation and interference rejection combining</td>
</tr>
<tr>
<td>Antenna gain</td>
<td>Macro: 14 dBi, SC: 5 dBi</td>
</tr>
<tr>
<td>Path loss</td>
<td>Macro cell: $140.7 + 36.7 \log_{10}(R[\text{km}])$ SC: $128.1 + 37.6 \log_{10}(R[\text{km}])$</td>
</tr>
<tr>
<td>Shadow fading</td>
<td>Lognormal, $\mu = 0$, std.=8 dB for macro cell</td>
</tr>
<tr>
<td></td>
<td>Lognormal, $\mu = 0$, std.=10 dB for SC</td>
</tr>
</tbody>
</table>

TheportionofsystemresourcesremainingaftersubtractingthereservedpartateachBS,iscalled sharedpartanditcanbeusedbyanyincommingtrafficflow.Accordingto[19],anoperatorrequiresatleast aminumportionofresourcestobereservedforitsuserswithinaBS,inordertoguaranteeQoS forparticulartrafficflows.Insimulations,forusersdownloadingFTPfilesitissetto50%[19]correspondingtotheschemenamedPRR50%. AlthoughsettingahighvalueforsharedpartwithinABScanleadtomoreflexibleallocationofresources,itiscomeswiththeshortcomingofnotcoveringthemminimumrequirementsforQoSoperators.However,weusethismaximumdegree offlexibilityinPRR,having0%RBreservedpartand100%sharedwithineachBS(i.e.,“PRR100%”),tocalculate thetheoreticalupperboundoftheaggregate throughput.

Security: Public
2.2.7.2 Network Performance

Figure 6 presents the aggregate system throughput (a metric indicated by 3GPP in [29][39]) with respect to an increasing offered traffic load for NVS as well as PRR50% and PRR100% with and without RENEV. As it may be observed, the experimental and theoretical curves for PRR100% and RENEV + PRR100% (the upper bound expressions as derived in Section V) match. For offered loads equal to 18 Mbps, the system’s behavior is the same for all the depicted schemes; all demanded traffic is served.

However, as the load increases, all compared schemes are able to serve less users compared to the system where RENEV is applied. In particular, when saturation is reached due to a lack of resources (i.e., offered load equals 78 Mbps), the throughput achieved with RENEV + PRR100% (60.93 Mbps) represents an increase of 50.68% with respect to PRR100%. In the first case, the available resources of other tiers are distributed according to traffic demand to cover the maximum number of users’ needs; however, when RENEV is not applied, each BS manages its own resources which are depleted after a while. At the other extreme, the NVS scheme achieves the poorest performance, since resources from different slices cannot be shared regardless of the traffic demands in each slice. The maximum value in this case is 23.19 Mbps. As for PRR50%, with and without RENEV, its performance constitutes an intermediate situation.
Notwithstanding the good results offered by PRR100% compared to PRR50% (both of them without the application of RENEV), the authors in [19] expound that a minimum share of the available resources should be reserved for each traffic slice to guarantee minimum QoS requirements. Therefore, PRR100% is not convenient in terms of QoS despite outperforming PRR50% in terms of aggregate throughput. The same conclusion applies when RENEV is implemented.

By inspecting Figure 6, it is particularly worth noting that RENEV + PRR50% (which as explained does not degrade the QoS requirements of the traffic slices) is even able to show higher aggregate throughput than PRR100%. This behavior is due to the ability of RENEV to compensate not only the traffic spatial non-uniformities but also the QoS loss experienced when sharing the 50% of the resources per BS, instead of the 100% in PRR.

**Figure 7. (a) Percentage of transferred RBs by each tier. (b) Traffic Served by each tier.**

Figures 7(a) and 7(b) study the percentage of transferred RBs per tier as well as the corresponding served traffic for the case of RENEV + PRR 100% (as depicted in Figure 6). As expected, the resources are provided to the tenant BSs first by the SCs tier and subsequently by the eNB. This observation is justified by the fact that the SCs are acting first as Donor BSs, thus their resources are first depleted; subsequently the role of the Donor BSs...
is undertaken by the eNB. In addition, we observe that the RB transfer first increases, then reaches a specific peak and then decreases for both tiers. The two peaks in Figure 7(a) equal 32.2% of transferred RBs by the SCs tier (achieved for 60 Mbps) and 32.64% by the eNB (achieved for 78 Mbps). After these peaks, although the number of users requiring resources is augmented, the transferred resources decrease because both tiers run out of RBs since all of them are already allocated to the existing users. Finally, it is worth commenting on the fact that the traffic being served by each tier (Fig. 5(b)) depends on the available number of RBs, as it can be observed in Fig. 4 as well. In particular, when the percentage of transferred RBs falls the aggregate throughput in Fig. 4 stabilizes since the resources are depleted and the incoming user requests cannot be satisfied. For the aforementioned peaks of RB transfer, 33.32 Mbps are served by the SCs and 34.2 Mbps by the eNB.

2.2.7.3 User's Throughput

In Figures 8(a), 8(b) and 8(c) we study the Cumulative Distribution Function (CDF) of user throughput (indicated metric in [29][39]) for three cases of traffic load: low offered load where the majority of users are served, medium one and that where the system is saturated; 42 Mbps, 66 Mbps and 78 Mbps correspondingly, as also depicted in Fig. 4. Here and subsequently let us focus on the case of PRR 100% with and without RENEVs since it provides the upper bound of network’s throughput.

First, we observe that the gains in throughput acquired in the network side with the application of RENEV can be translated into merits for the end users. According to Fig. 8(a), as the offered load is slow, RENEV is able to help the majority of users to achieve the demanded data rate. In particular, the observed slight deviation from 300 kbps is due to the fact that some users do not achieve the demanded data rate because of the channel conditions they experience. However, without applying RENEV the user throughput dispersion is quite high. For instance, 80% of the users achieve throughput values equal to or higher than 250 kbps. The rest 20% of the users achieve values ranging from 120 kbps to 250 kbps.
In addition, we observe that a higher offered load affects dramatically the user throughput. For example, in Fig. 8(b), 72% of the users achieve a transmission rate equal or higher than 250 kbps when RENEV is applied. On the other hand, for the same percentage without applying RENEV, the lowest user throughput value achieved is 130 kbps. In particular, the transfer of resources defined by RENEV improves the performance of users with poor links, who are normally located in the cell edge area. These users are more demanding in terms of required RBs in order to be served. However, RENEV is able to satisfy such kinds of users. For instance, when the system is further loaded (Fig. 8(c)), the dispersion among user throughput values is quite high, both with and without the application of RENEV. Even in this study case, 50% of the overall users achieve 75% of the demanded transmission rate (with lowest user throughput equal to 102 kbps). On the contrary, without RENEV, this percentage falls to 52.5% of the demanded data rate.

### 2.2.7.4 Signaling Overhead

In this set of our experiments, we evaluate the requests and the corresponding messages that are necessary for the transition from a scenario where all resources are initially distributed uniformly among the BSs, to a scenario where the resources are finally distributed according to the existing geographical traffic variations.
In Figure 9 we study the impact of the number of SCs into the percentage of successful requests per cluster, for different traffic offered loads (i.e., low, medium, and high as in Figure 8). It is worth commenting on the fact that in dense scenarios in terms of SCs, the available RBs are quickly depleted, and therefore, the number of successful requests falls. This means that the tenant Requesting BSs cannot acquire the demanded resources. Moreover, for high loaded systems less requests are satisfied since resources are exhausted faster. For example, if a cluster with 6 SCs is considered (the same scenario analyzed in Figure 6), the percentage of successful requests is 86.5% for 42 Mbps offered load, 80% for 66 Mbps and 72% for 78% Mbps. On the other hand, when 10 SCs are considered within the cluster’s surface, this percentage falls to 77%, 70% and 61% for the corresponding cases.

![Figure 9. Percentage of successful requests for different number of SCs per cluster.](image)

Figure 10 studies the number of exchanged messages per SC, for the three offered loads studied so far. In all cases, the experimental results clearly showcase that higher number of SCs within the cluster’s surface, is translated into higher absolute number of exchanged messages over X2 interface. For instance, for a cluster with 6 SCs, we observe in average 8.5 exchanged messages for 42 Mbps, 10.4 for 66 Mbps and 12.4 for 78 Mbps. The last case refers to the saturation of the system, as presented in Figure 6. In particular, as the number of SCs in a cluster increases, the messages among the participant tenant BSs are also increasing even though the rate of increase progressively reduces.
The physical implementation of X2 is still not standardized, so it should be noted that it is the main factor imposing feasibility constraints. In general, we note that a particular number of SCs where RENEV can be applied depends on the limits inserted in the actual implementation of X2 and the corresponding capacity reserved for signaling. Figure 10 can result quite useful for operators, to calculate the actual signaling for a certain number of SCs per cluster, according to the way they choose to implement X2 (i.e., such as fiber, over-the-air wireless, etc.).

![Figure 10. Number of exchanged X2 messages per SC.](image)

### 2.2.8 Concluding Remarks

In this section, we have proposed RENEV, a scheme that considers the coordination among several BSs to create an abstraction of systems’ radio resources, so that multiple tenants (i.e., BSs) can be served in a heterogeneous environment. The extensive performance assessment has revealed that gains in the system’s throughput are translated into gains for the users’ throughput as well. With the use of RENEV, system’s resources are dynamically distributed according to users’ needs on an isolated and on-demand basis. In this way, the majority of the users is served, as long as spare resources exist. Finally, the solution has been evaluated for the signaling overhead that adds into the network for increasing the number of SCs per cluster. The proposed analytical formulation can provide useful insights that can be exploited for effective network planning of SC clusters in realistic heterogeneous environments.
2.3 Appendices

2.3.1.1 Appendix A: MCS Selection Probability

Let us denote by $x_i \in \mathbb{R}^2$ the location of $B_S_i$ and $y \in \mathbb{R}^2$ a random location in the scenario. The signal strength received from $B_S_i$ at location $y$, expressed in dB, may be written as $p_i(y) = P_T i - L_i(y) - S_i(y)$, where $P_T i$ includes antenna gains and transmitted power of $B_S_i$, $L_i(y)$ is the path loss from $x_i$ to $y$, and $S_i(y)$ is the slow fading. In general, shadowing is commonly described as a log-normal distributed random variable with zero mean and a standard deviation $\sigma_S$ typically around 8 dB for the eNB and 10 dB for the SC [39].

According to the notation, the SINR received at $y$ from $B_S_i$, when no interference is received, is given by $\text{SINR}_i(y) dB = p_i(y) - N_0$, where $N_0$ represents the noise average power. Throughout the analysis, taking into account the transmission power and coverage area of each $B_S$ as well as that subcarriers are not utilized by neighboring cells, we assume that interference is imperceptible among them [34][39]. For the sake of simplicity and clarity, and without loss of generality, the dependency of the several variables on the location $y$ will be omitted in these sequel. Yet, all expressions are still derived for a random location $y$. Therefore, let $\text{SINR}_{max}$ be the highest SINR received from a $B_S$ in a random location $y$, where $\text{SINR}_{max} = \max_{B_S_i \in B_S} \text{SINR}_i$.

Focusing on the adaptive MCS mechanism, the decision on the particular MCS to be used by a particular $B_S$ with a user located at $y$ is based on the received SINR $i$, i.e., $\text{SINR}_i$. Thus, in practice, the $k^{th}$ MCS is selected by $B_S_i$ if and only if $\text{SINR}_{min} \leq \text{SINR}_i < \text{SINR}_{max}$, where $\text{SINR}_{min}$ and $\text{SINR}_{max}$ stand for the minimum and maximum thresholds of MCS $k$, respectively. Therefore, the probability of using a certain MCS could be expressed as:

$$P(MCS_i = k) = P(\text{SINR}_k^{min} \leq \text{SINR}_i < \text{SINR}_k^{max} | \text{SINR}_i = \text{SINR}_{max})$$

$$= \frac{P(\text{SINR}_k^{min} \leq \text{SINR}_i < \text{SINR}_k^{max} \cap \text{SINR}_i = \text{SINR}_{max}) - P(\text{SINR}_i = \text{SINR}_{max})}{P(\text{SINR}_i = \text{SINR}_{max})}. \quad (18)$$

Since the SINR of a particular $B_S_i$ is considered independent from the SINR of the rest $B_S$s, the denominator of (18) equals
\[ P(\text{SINR}_i = \text{SINR}_{\text{max}}) = P(\bigcap_{j \neq i} \text{SINR}_j > \text{SINR}_j) = \prod_{j \neq i} P(\text{SINR}_j > \text{SINR}_j) \]
\[ = \prod_{j \neq i} P(P_{T_j} - L_i - S_i > P_{T_j} - L_j - S_j) \]
\[ = \prod_{j \neq i} P(S_j > S_i + \mu_{ij}), \quad (19) \]

where \( \mu_{ij} = P_{T_j} - P_{T_i} + L_i - L_j \). Based on the analysis provided in [15] and after a convenient change of variables, (19) is equal to \( F_{S_i}(\frac{\sigma_{ij}}{\sigma_0}) \), where \( F_{S_i} \) denotes the Cumulative Distribution Function (CDF) of the random variable \( S_i \) expressing the shadowing, whereas \( \sigma_i \) and \( \sigma_j \) denote the standard deviations of the shadowing for the participating pair of BSs (i.e., for two SCs or for a pair SC-eNB).

Correspondingly, the numerator of (18) is derived in [15] as
\[ P(S_{\text{min}}^0 \leq S_i < S_{\text{max}}^0 \cap \text{SINR}_i = \text{SINR}_{\text{max}}) \]
\[ = \prod_{j \neq i} P(S_{\text{min}}^0 \leq S_i < S_{\text{max}}^0 \cap \text{SINR}_i > \text{SINR}_j), \quad (20) \]

since the probability of joint occurrence of pairwise independent event equals the product of the events’ marginal probabilities. By substituting the values \( S_{\text{0}} = P_T - S\text{INR}_{\text{max}}^k - L_i \) and \( S_{\text{1}} = P_T - S\text{INR}_{\text{min}}^k - L_i \), (20) is expressed as follows:
\[ P(S_{\text{0}} \leq S_i < S_{\text{1}} \cap S_j > S_i + \mu_{ij}) = \int_{S_{\text{0}}}^{S_{\text{1}}} \int_{S_i + \mu_{ij}}^{\infty} f_{s_j}(s_j)f_{s_i}(s_i)ds_jds_i \]
\[ = \left( F_{s_i}(S_{\text{1}}) - F_{s_i}(S_{\text{0}}) \right) - \int_{S_{\text{0}}}^{S_{\text{1}}} F_{s_i}(s_i + \mu_{ij})f_{s_i}(s_i)ds_i. \quad (21) \]

According to the analysis presented in [15] this expression also depends on the Probability Distribution Function (PDF) (i.e., \( f_{s_i} \)) and CDF (i.e., \( F_{s_i} \)) of the random variables for the shadowing and it is calculated numerically.

2.3.1.2 Appendix B: Derivation Of Equation (6)

For deriving the throughput achieved by the users located within the SCs tier with RENEV, let us divide the process according to the source that provides RBs to the requesting BSs. First, resources are redistributed within the SCs tier to serve the demanded traffic. In the case that these are not enough, resources are granted from the eNB.

To begin with, SCs tier redistributes its RBs to accommodate the demanded traffic. If the overall traffic is less or equal to the SCs capacity, all users can be served. The overall resources within this tier are equal to \( RB_T = \sum_{i \neq 0} RB_i \). What is more, the average
transmission rate for this case, equals \( E[R_{TOT}] \equiv \frac{1}{1-a_0} \sum_{i \neq 0} a_i E[R_i] \), where \( a_i \) denotes percentage of users located within the coverage area of \( BS_i \). Thus if \( \sum_{i \neq 0} X_i d \leq RB_T E[R_{TOT}] \), all users located in the SC tier will be served by the SC’s resources. It follows that \( \sum_{i \neq 0} T_{R_i} = d \sum_{i \neq 0} X_i I_i \).

Once SCs’ resources (i.e., \( RB_T \)) are depleted, \( X \cdot (1 - a_0) \) users within the Requesting BSs, will require further resources from the eNB tier. Therefore, the expected number of users to be served with resources from the eNB is

\[
E = X \cdot (1-a_0) - \frac{RB_T}{d} \cdot E[R_{TOT}].
\]

Thus, each Requesting BS will have to serve

\[
E_i = \frac{a_i}{1-a_0} \cdot E
\]

users. Let us denote as \( E[R_{BC}] \) the amount of resources from the eNB that can be given to each Requesting BS. Based on this, a particular Requesting BS will serve all traffic in the case where the following holds:

\[
E_i \cdot d \leq E[R_{BC}] \cdot E[R_i].
\]

In contrast, the traffic served by Requesting BSs with RBs from the eNB tier will be equal to \( E[R_{BC}] \cdot E[R_i] \). Consequently the throughput generated by the users in the SCs tier, served both with resources redistributed within the SCs tier and resources transferred from the eNB, will be equal to

\[
\sum_{i \neq 0} T_{R_i} = \min \left( X \cdot (1-a_0) \cdot d, RB_T \cdot E[R_{TOT}] + \sum_{i \neq 0} E[R_{BC}] \cdot E[R_i] \right)
\] (22)

which corresponds to (5) included in Section 2.2.5.

It remains to show how we calculate the number of eNB resources that can be given to each Requesting BS (i.e., \( E[R_{BS}] \)), included in (22). For the sake of simplicity and without loss of generality, we can assume a circular cluster’s surface containing \( N \) circular shaped SCs. Therefore, the area of the cluster is \( A = \pi R^2_c \) where \( R_c \) is the cluster radius, and \( A_i = \pi R^2_{SC_i} \) and \( A_j = \pi R^2_{SC_j} \) hold for two random circular shaped SCs’ coverage areas, having radius \( R_{SC_i} \) and \( R_{SC_j} \) respectively. For a particular Requesting BS, located randomly within the cluster, there will be an overlap if the distance between \( BS_i \) and
another $BS_j$ is less than the summation of their radii. Thus the probability of overlap among two Requesting BSs is derived as

$$P_o = \frac{\pi \cdot (R_{SC_i} + R_{SC_j})^2}{\pi \cdot R_c^2} = \left(\frac{R_{SC_i} + R_{SC_j}}{R_c}\right)^2. \quad (23)$$

Then, the probability for a Requesting $BS_i$ of having $n_i$ overlaps is described by a binomial random variable as follows:

$$P(n_i = n) = \binom{N - 1}{n} \cdot P_o^n \cdot (1 - P_o)^{N-1-n}. \quad (24)$$

We assume that a Requesting $BS_i$ with $n_i$ overlapping BSs, receives \( \frac{RB_s}{n_i+1} \) RBs. The expected value of this term is equal to

$$E\left[\frac{RB_s}{n_i+1}\right] = \frac{RB_s}{P_o} \cdot \sum_{m=1}^{N} \frac{(N - 1)!}{m!(N - m)!} \cdot P_o^m \cdot (1 - P_o)^{N-m}, \quad (25)$$

where a convenient change of variables can be applied $m = n_i + 1$, so as (25) equals

$$E\left[\frac{RB_s}{n_i+1}\right] = \frac{RB_s}{P_o} \cdot \sum_{m=1}^{N} \frac{N!}{m!(N - m)!} \cdot P_o^m \cdot (1 - P_o)^{N-m} = \frac{RB_s}{NP_o} \cdot [1 - (1 - P_o)^N],$$

which is valid for all Requesting BSs since each one is assumed to receive $\frac{RB_s}{NP_o}$. However this is not true in the case that each Requesting BS $BS_i$ accommodates different portion of users (i.e., $a_i$). This difference in the Requesting BS load, implies different traffic demands and hence unequal percentage of resources to be allocated.

Let us assume as previously, that Requesting $BS_i$ overlaps with $n_i$ BSs. The number of resources that would be allocated per Requesting $BS_i$ from the eNB can be expressed as

$$E[RB_i^x] = RB_s \cdot \sum_{n_i=0}^{N-1} P_o^{n_i} \cdot (1 - P_o)^{N-1-n_i} \cdot \sum_{c=1}^{\binom{N-1}{n_i}} \frac{a_i}{\sum_{BS_k \in C_i} a_k}. \quad (27)$$
It should be noticed that for equal percentage of users in each Requesting BS (i.e., when $a_i = a_k \forall i \neq k$), (27) is equal to (26).

### 2.3.1.3 Appendix C: Set Of Feasible Future States

RENEV is intended to redistribute the unused resources of the possible Donor BSs among the Requesting BSs. Therefore, not all transitions from state $S_n$ to state $S_j$ are feasible. The set of feasible future states for a given state $S_n$, $F(S_n)$, is defined as the set of states to which $S_n$ could transit after performing RENEV. Therefore, the cardinality of the set $F(S_n)$ is derived according to how many of the following conditions are accomplished. Then, we derive the transition probability $p_{nj}$ as defined in (13). Hence, and based on the definition of states $S_n$, $S_j$, and RENEV algorithm, the following conditions must be accomplished to assure that $S_j \in F(S_n)$:

- The amount of resources is constant in the initial and the final states:
  \[
  \sum_{k=1}^{r_{\text{max}}-r_{\text{min}}+1} s_{j,k} \cdot (r_{\text{min}} - 1 + k) = \sum_{k=1}^{r_{\text{max}}-r_{\text{min}}+1} s_{n,k} \cdot (r_{\text{min}} - 1 + k) \tag{28}
  \]

- After performing RENEV, the number of Requesting BSs should be smaller:
  \[
  n_R(S_j) < n_R(S_n) \tag{29}
  \]

- The number of requested RBs in the final state should be less than the corresponding number in the initial state:
  \[
  \sum_{k=1}^{-r_{\text{min}}} s_{j,k} \cdot (r_{\text{min}} - 1 + k) < \sum_{k=1}^{-r_{\text{min}}} s_{n,k} \cdot (r_{\text{min}} - 1 + k) \tag{30}
  \]

- In state $S_j$ (i.e., final state) there are not new Requesting BSs:
  \[
  \forall s_{n,k} = 0, k \leq -r_{\text{min}} \Rightarrow s_{j,k} = 0 \tag{31}
  \]
  \[
  \forall s_{n,k} \neq 0, k \leq -r_{\text{min}} \Rightarrow s_{j,k} \leq s_{n,k} \tag{32}
  \]

- The number of RBs transferred by the Donor BSs is equal to the number of RBs received by the Requesting BSs:
  \[
  \sum_{k=2}^{r_{\text{max}}-r_{\text{min}}+1} (s_{n,k} - s_{j,k}) \cdot (r_{\text{min}} - 1 + k) = \sum_{k=1}^{-r_{\text{min}}} (s_{n,k} - s_{j,k}) \cdot (-r_{\text{min}} + 1 - k) \tag{33}
  \]
The absolute value of the highest amount of requested RBs in the initial state (i.e., negative value) should be lower or equal than the minimum amount of available RBs such that:

\[
|k'| = r_{\max} - r_{\min} + 1, \quad \forall k \leq -r_{\min}, \quad \forall s_{n,k} \neq 0 \quad \text{then} \quad s_{jk} \neq 0, \quad \forall k \geq |r_{\min} - 1 + k'| \quad (34)
\]

As RENEV is completed, all possible redistribution of resources has been done. Therefore, there is not any possible Donor BS that could cover the needs of a Requesting BS:

\[
\forall s_{j,k} \neq 0, k \leq -r_{\min} = \sum_{m=-2r_{\min}+2-k}^{r_{\max}-r_{\min}+1} s_{j,m} = 0 \quad (35)
\]

### 2.3.1.4 Appendix D: Derivation of \( P (Q = q|N,M) \)

The eNB coverage area is overlaid with SCs clusters, each one containing \( N \) SCs. Resources can be allocated simultaneously to more than one non-overlapping Requesting BSs, when eNB undertakes the role of Donor BS. Throughout the deployment area, we assume clusters that do not overlap among each other, so without restriction of the general case we induce our analysis to one cluster of SCs.

The probability that two BSs within the cluster are overlapping is derived in (23), denoted as \( P_o \), and the probability that a specific BS \( i \) in the cluster is overlapped with \( n \) BSs, denoted as \( P (n_i = n) \), is derived in (24). Note that, for a given state \( S_j \), if BS \( i \) is assumed to be a Requesting BS, the probability that a BS different from BS \( i \) is a Requesting BS equals \( P_{(N,M)}^{m-1, n-1} \), where \( M = n_R(S_j) \).

Let us denote with \( m_i \) the number of Requesting BSs overlapping BS \( i \). Henceforth, \( P_{RB}(m = m\mid N,M) \) denotes the probability that \( m \) Requesting BSs overlap BS \( i \), given that \( M \) out of \( N \) BSs are Requesting BSs, and can be expressed as:

\[
P_{RB}(m_i = m\mid N,M) = \sum_{k=m}^{N-1} \left( P(n_i = k) \cdot \binom{k}{m} \cdot P_{N,M}^m \cdot (1 - P_{N,M})^{k-m} \right). \quad (36)
\]

In RENEV, the eNB will only transfer the same resources to two different Requesting BSs if they do not overlap. Approximately, we could claim that the available resources of the eNB can be transferred to a specific SCs cluster, as many times as the number of non-overlapping groups of Requesting BSs. Therefore, we are interested in figuring out the
number of non-overlapping groups of Requesting BSs within the cluster, denoted as \( Q = \{1, 2, \cdots, M\} \). For instance, when all Requesting BSs overlap altogether, \( Q = 1 \); when there are two non-overlapping groups of BSs, \( Q = 2 \) (i.e., BSs are overlapped within each group but non-overlapped with the BSs of the other group); finally, when all Requesting BSs are not overlapped, \( Q = M \).

If we assume that all BSs within each group overlap with each other, the probability of having \( Q \) non-overlapping groups of BSs can be approximated by

\[
\begin{align*}
P(Q = q|N, M) &
= \left\{
\begin{array}{ll}
P_{RB}(m_i = N - 1|N, M) & \text{if } q = 1, \\
\sum_{k=0}^{M-1} P_{RB}(m_i = k|N, M) \cdot P_{RB}(m_i = M - q - k|N, M)) & \text{if } q = 2, \\
\sum_{k=0}^{M-q} P_{RB}(m_i = k|N, M) \cdot P(Q = q - 1|N - 1 - k, M - 1 - k) & \text{if } q > 2.
\end{array}
\right.
\end{align*}
\]

3. **Open eNB (UoA)**

The conventional business model followed by the network operators, based on a dedicated network infrastructure paradigm where each operator have access to a fixed set of network resources, networking equipment, hinders the employment of flexible and spectrum-agile resource management, further increasing the capital and operational expenses required for the mobile network operation, i.e. CAPEX and OPEX, respectively. In order to improve the cost performance of their networks, the operators are required to maximize the effectiveness of the resource utilization stage, e.g. by enabling more efficient sharing of physical network components including base stations or core network elements. Moreover, operators need to adopt efficient capacity boosting solutions to meet the growing demand of the mobile traffic. In this direction, network densification is a straightforward solution for boosting the area spectral efficiency [40]. However it comes at the cost of increased monetary and network management overheads due to the increased number of small-sized base stations involved. Nevertheless, combining network densification with a flexible sharing of the Radio Access Network (RAN) among different operators enables more flexibility in the resource management and a consequently reduction of the CAPEX. In this context, the aim of this research activity is to devise architectural innovations for enabling virtually increasing of the base station density of a mobile network operator by promoting on-the-fly leasing of the physical networking infrastructure from other mobile network operators.
Regarding the RAN sharing, 3GPP has recently outlined a suite of requirements and guidelines for sharing the access network among multiple network operators [11]. Nevertheless, the detailed implementation of the resource sharing mechanism is left open. In this direction, the authors in [22] propose a scheme for capacity sharing between different operators. A simple model for base station sharing in LTE is also presented based on creating a group of logically independent virtual base stations, referred to as Virtual eNodeBs (VeNBs). A key feature of the VeNBs, is that they can be operated by different operators at the same time. Although simulation results demonstrate significant improvements in terms of load balancing at the base stations, more light needs to be shed on the adaptability, flexibility, and elasticity of the proposed model in the future fifth generation (5G) network.

In another work a procedure for resource allocation management in a multi-tenant base station is proposed [41]. This procedure enables operators to deploy independent resource scheduling policies among virtual eNBs instances running in a shared LTE base station. To achieve this, a dynamic two-layer resource scheduler is employed. This scheduler is composed by i) a common physical scheduler, which is responsible for frequency domain scheduling, and ii) a set of virtual schedulers, each of which gives the VeNBs the freedom to implement customizable scheduling policies. In a similar work in [42], the authors focus on the MAC functions required for allocating the resources within a shared evolved NodeB (eNB) among multi-tenant operators. To achieve this, they introduce an entity called Hypervisor that is responsible for allocating the available physical resource blocks (PRBs) among multiple operators in line with predefined service level agreements (SLA).

Different from [22],[41],[42] this research activity promotes the use of network virtualization (NV) [43] and SDN-based architectural innovations [44] for enabling a better coordination among multi-tenant base stations and increasing network capacity by exploiting the underutilized resources in existing base stations. In this direction, an SDN-based network architecture is proposed where all the decisions regarding the multi-tenancy process are moved in centralized logically unity, referred to as SDN Controller, that maintains a global knowledge of the network state. The proposed SDN-based network architecture and the functional and architectural enhancements introduced by the proposed solution are described in Section 3.1. Section3.2 provides a comprehensive discussion on the signaling flow required to support the proposed solution in the LTE/LTE-A system, while Section 3.3 presents extensive system-level simulations to assess the performance of the proposed solution.
3.1 Proposed Architecture

Figure 11 depicts the proposed multi-tenant RAN architecture for dynamic base station sharing among multiple LTE/LTE-A network operators. As in the baseline version of LTE/LTE-A, the evolved eNBs provide user and control plane protocol terminations towards the UE. Moreover, the LTE-A Core Network (CN) is composed by mobility management entities (MMEs), serving gateways (S-GWs), and a single Packet Gateway (P-GW). The MME is responsible for handling the control plane functions for mobility management throughout the mobile operator’s network. On the other hand, the S-GW handles the user plane, as it is responsible for routing and forwarding the user data packets towards the P-GW. The P-GW is responsible for inter-connecting the remainder RAN elements to the Internet.

Apart from the aforementioned architectural elements, the proposed architecture includes an additional entity, referred to as Main Controller (MC), which lies in the LTE/LTE-A CN. The MC is a software unit that maintains global knowledge of the status of the mobile operator’s network in an SDN fashion. Without loss of generality, it is considered the presence of two mobile network operators (Figure 11), each one of which has its own RAN/CN equipment, respectively. Furthermore, it is assumed that a tagged UE is registered in the operator A’s network (UE A) and is currently served by a legacy eNB: the eNB A. It is assumed that that operator B utilizes a set of base stations (BSs) endowed with NV functionalities. This type of base stations will be referred to as Open eNBs (OpeNBs). The OpeNBs are assumed capable of sharing their physical resources among multiple operators in a dynamic fashion. The OpeNB research idea has been introduced in [45] and described in the CROSSFIRE deliverable D2.1 as well. The sharing process of the OpeNB is based on SLAs that are established between the involved operators (operators A and B in Figure 11), prior to the sharing of the physical resources.
In Figure 11, it is assumed that operator A can lease the base stations owned by operator B, i.e. the multi-tenant OpeNBs, and perform handovers (HOs) towards them if a certain set of conditions are satisfied. Apart from employing traffic offloading, or attaining an improved UE performance, such functionality enables the cellular UEs to connect to the closest cellular BS as well (independent of the operator). Aiming to clarify the role of the additional entities present in the proposed architecture, a description of the key functionality of the OpeNBs and the MC at the different mobile operators is provided in the sequel. Moreover some architectural modifications needed to support logic connectivity are discussed, e.g. tunneling, between the OpeNBs of the host operator (Operator B) and the CN of the home operator (Operator A).

### 3.1.1 OpeNB

An OpeNB is capable of supporting all functions supported by a legacy eNB, but also acts as a multi-tenant base station. To achieve this, the OpeNBs adopt a slightly modified protocol stack compared to the one supported by the legacy eNBs. In particular, the physical (PHY) layer of an OpeNB provides similar functionality with the one provided by the PHY of a legacy eNB. Nevertheless, an additional layer, termed as the Hypervisor Layer, runs on the
top of the PHY layer. The purpose of this layer is to abstract the PHY from the upper layers of the protocol stack. To achieve this, the Hypervisor layer virtualizes the physical resources and allocates them to a pool of virtual software instances, termed as virtual eNBs (VeNBs). Each VeNB emulates the behavior of the remainder upper layers of the protocol stack, while it is logically connected to the CN of the appropriate operator by means of a Layer 3 (L3) tunnel link. In more detail, the Hypervisor Layer is considered to employ the Network Functions Virtualisation (NFV) technology, towards achieving the virtualization of the physical base station infrastructure. At this point, it is worth to note that a typical NFV-based architecture is composed by numerous building blocks of virtualized network functions (VNFs) that run on top of the physical network infrastructure. In the OpeNB architecture, these VNFs are software instances that are responsible for handling specific network functions, including the functionalities of the upper layers in the eNB protocol stack. The VNFs are considered to be chained together in a building block fashion, with the aim to deliver a full networking service, e.g. to emulate the behavior of each VeNB. Besides, the Hypervisor Layer handles the logic interface between the PHY layer functions and the VNFs that emulate the upper layers, by means of NFV application programming interfaces (NFV APIs). In such operations, the Hypervisor Layer is assisted by the MC that lies within the CN of the host operator, e.g. the MC-Host in Figure 11.

3.1.2 Main Controller

The MC is a (logically) centralized entity that is considered capable of acquiring a global knowledge of the network state by exchanging information with all base stations of the mobile operator’s network. To this end, a minor software upgrade is required at the legacy base stations to let them maintain a logic communication channel with the MC. In this channel, the base stations provide some information about their network status, e.g. traffic load, PHY measurements of their served UE, or whatever information requested by the MC. The MC, which lies at the CN of the home operator (MC-Home), exploits the acquired knowledge to perform mobility management operations in a centralized fashion. On the other hand, the MC of the host operator (MC-Host) utilizes its knowledge to handle the resource sharing operation at the OpeNBs. More specifically, the MC-Home, i.e. the MC-Home in Figure 11, is responsible for taking the HO decisions towards an OpeNB of the host operator. In the sequel, such decisions are termed as NV Decisions. The functional migration of the
HO decision to the CN is in line with the C-RAN architecture and the SDN paradigm, while it also enables the operator to better handle user mobility in a centralized fashion. Note that during a NV Decision, the MC-Home identifies the most appropriate target base station among the OpeNBs of the host operator (see Section 3.2.2 for more details). To this end, the MC-Home maintains a logic channel with the MC-Host to exchange information about the status of the OpeNBs of the host operator, in a real-time fashion. This channel is also used by the two MCs to establish a tunnel connection between the OpeNB of the host operator and a CN entity of the home operator. The tunnel establishment phase is discussed later. Conversely, the MC-Host manages the multi-tenant operation at the OpeNBs and provides all necessary functionality to handle the sharing of resources among multiple VeNBs. More specifically, the MC-Host assists the Hypervisor Layer at the OpeNBs in configuring the logic connection between the UEs served by an OpeNB and the appropriate VeNB instance. It is also responsible for dynamically creating the VeNBs instances in each OpeNB and monitoring the leasing conditions established in the SLA agreements. Finally, the MC-Host configures the logical connection between each VeNB instance and the respective CN of the home operator.

3.1.3 L3 tunnel

Aiming to clarify the role of L3 tunnel, in the sequel it is assumed that the UE A of the home operator is served by the OpeNB of the host operator (Figure 11). To achieve this, the VeNB should be logically inter-connected to the CN of the home operator, e.g. by means of an L3 tunnel. Such a tunnel is required for providing a logical interface between the control/data plane protocols running, on the one hand, at the VeNB instance at the OpeNB, and, on the other hand, at an appropriate attachment point in the CN of the home operator, e.g. the MME/S-GW of the serving eNB. Since this procedure should be transparent to the CN entities of the home operator, it is assumed that the tunnel establishment and maintenance is performed by the MC-Home, which is additionally responsible for forwarding the control/data plane packets from the MME/S-GW at the home operator to the target VeNB, in an OpenFlow fashion [44]. OpenFlow is a communication protocol that enables a remote controller to program the forwarding table of an SDN capable switch that is typically referred to as the OpenFlow Switch. OpenFlow is used in the proposed architecture, as it permits a more effective maintenance of the L3 tunnel. In more detail, it provides the MC-
Home, which acts as a remote controller, with the opportunity to update, remove or create on-the-fly new logic interfaces with the OpeNBs of the host operator. However, in order to enable the usage of OpenFlow at the CN, a slight architectural modification is needed at the MME/S-GW. In particular, it is assumed that the OpenFlow Switch is used to perform all forwarding operations that involve the MME/S-GW. Also, it is considered that the control plane functionalities, i.e. the ones required for executing the OpenFlow operations, are moved to the MC-Home. The use of an OpenFlow Switch to the CN is considered as a non-invasive architectural enhancement, since it can provide transparent connectivity between the entities that reside in the CN, e.g. the eNB and the P-GW.

3.2 OpeNB Sharing Procedure

Figure 12 shows the signaling flow for implementing the proposed NV-based multi-tenant operation of the RAN. The signaling flow below is performed for the first time when a UE registered to Operator A is transferred to an OpeNB that is owned by the Operator B. The first two steps are integral part of the baseline HO procedure. The serving eNB sends to the tagged UE a “Measurement Configuration” message, specifying the set of physical measurements that need to be performed (step 1). In the following, the UE reports to the serving eNB the results of the respective measurements (step 2) and, based on the reported measurements, the serving eNB decides whether a HO towards an OpeNB is required or not (step 3). This type of HO is termed as NV-HO in the sequel. The triggering of a NV-HO can be based on different types of criteria that relate with the current network status and can be re-programmed at the eNBs by the corresponding MC (MC-Home).

Some exemplary NV triggering algorithms based on the signal quality at the UE or the cell load at the eNBs of Operator A are discussed in the next Section. If the triggering criteria for a NV-HO are met, the serving eNB forwards all necessary information on the respective UE (including the derived measurements) to the MC-Home (step 4). Accordingly, based on its global knowledge on the network status, the MC-Home identifies the OpeNB of the host operator (operator B) that satisfies a predefined set of NV decision criteria (step 5). This phase is referred to as the NV Decision phase in the sequel. Once the NV decision is taken, the MC-Home sends an OpeNB Access Request to the MC-Host (step 6), containing the identity of the home operator (Operator A), the identity of the target OpeNB, as well as any information related to the characteristics of the ongoing UE connections.
The identity of the home operator is required at the host operator (operator B) to verify whether the NV-HO request is in line with SLA agreements between the two operators (step 7). On the other hand, the identity of the target OpeNB is needed at the MC-Host to forward the OpeNB Access Request towards the respective OpeNB (step 8). Upon receiving the OpeNB Access Request message, the target OpeNB utilizes the information on the characteristics of the ongoing UE connections (included in step 6) in order to infer on

Figure 12: OpeNB Sharing Signaling Flow

The identity of the home operator is required at the host operator (operator B) to verify whether the NV-HO request is in line with SLA agreements between the two operators (step 7). On the other hand, the identity of the target OpeNB is needed at the MC-Host to forward the OpeNB Access Request towards the respective OpeNB (step 8). Upon receiving the OpeNB Access Request message, the target OpeNB utilizes the information on the characteristics of the ongoing UE connections (included in step 6) in order to infer on
whether it can resourcefully host the connections of respective UE (step 9). This phase, referred to as the NV Admission Control phase, can be based on regular call admission control schemes that can be found in current literature [14].

In case of successful admission, the OpeNB notifies the MC-Host that it is capable of hosting the connections of the respective UE, by sending an “OpeNB Access Ack” message (step 10). Accordingly, upon receiving this ack message from the OpeNB, the MC-Host confirms the admission of the tagged UE to the target OpeNB by replying with a VeNB instance request message for Operator A (home operator) (step 11). In the sequel, the OpeNB reserves an appropriate set of resources for the tagged UE, creates a VeNB instance for Operator A locally, and confirms both actions to the MC-Host (step 12).

Upon receiving the confirmation message from the OpeNB, the MC-Host requests the establishment of an L3 tunnel between i) the VeNB instance at the target OpeNB and ii) a CN entity of Operator A, and confirms the reservation of resources to the home operator (Operator A) (step 13-15). Accordingly, the MC-Home sends an L3 tunnel configuration request to the MME/S-GW of the current serving eNB (step 16) and the respective MME/S-GW proceeds with the establishment of a tunnel towards the target OpeNB, i.e. the respective VeNB instance (step 17). Note that the purpose of the L3 tunnel is to interconnect the VeNB at the host operator to the CN of the home operator, as it would have been a legacy eNB of the home operator. Having attached the VeNB to the MME of the serving eNB at the home operator, the VeNB forwards a NV-HO acknowledgement message towards the serving eNB through the L3 tunnel established with their common MME (steps 18 and 19). Upon reception of the NV-HO Acknowledgement, the serving eNB sends a HO command to the UE (step 20) and the UE performs legacy handover execution procedures towards the target OpeNB (step 21). Due to the presence of the VeNB instance at the OpeNB, the UE perceives the VeNB at the OpeNB as a legacy eNB that belongs to the home operator (operator A). A more detailed description of the NV phases (see the yellow boxes in Figure 12) is presented below.

### 3.2.1 NV-Triggering Phase

The NV Triggering phase is executed in step 3 (Figure 12). Firstly, it is worth to note that HO triggering is integral part of the legacy HO execution procedure in LTE/LTE-A as well, where the received signal strength at the UE, the received signal quality at the UE, or the cell load, are compared to absolute or relative thresholds. The key difference of the NV
Triggering phase in the propose architecture, as compared to the legacy one, is that the set of candidate eNBs is enriched with the set of OpeNBs belonging to different operators. To this end, in this phase, both types of HOs are triggered: the ones towards a base station of the operator A (as in the legacy procedure) and the ones towards the OpeNBs of another operator. The choice of the NV triggering algorithm should be left open to the network operators, to enable them employ their own network management strategies. In the following, two simple exemplary NV triggering algorithms: the NV-RSRQ and the NV-Offloading algorithm are proposed.

In the NV-RSRQ algorithm, the serving eNB triggers a (NV-)HO based on the Reference Signal Received Quality (RSRQ) measurements that are performed by the tagged UE and reported periodically in the serving eNB (steps 1-2). In more detail, the UE reports the RSRQ of all base stations in proximity. If the base station with the highest RSRQ belongs to the home operator, the serving eNB performs legacy handover execution. On the contrary, if the base station with the highest RSRQ belongs to another operator (OpeNB), the serving eNB sends a NV-HO request to the MC-Home and includes all reported RSRQ measurements by the UE. At this point, it is worth to note that this algorithm can be employed by re-configuring the UE to perform RSRQ measurements in bands other than the ones of the home operator. This is possible in LTE/LTE-A since, on the one hand, the serving eNB indicates to the UE a list of frequencies for cell search [46] and, on the other hand, the UEs are capable of scanning and performing measurements in all bands available for cellular communications.

In the NV-Offloading algorithm, the serving eNB triggers a (NV-)HO based on its cell load. In more detail, if the current cell load of the serving eNB is lower than a prescribed threshold, the serving eNB handles the execution of HOs as in the legacy HO execution phase. However, if the cell load of the serving eNB is higher than a threshold, the serving eNB triggers the UE to perform RSRQ measurements for all LTE-A base stations in proximity. Accordingly, the serving eNB follows a similar procedure with the NV-RSRQ algorithm. In more detail, if the base station with the highest RSRQ belongs to the home operator, the serving eNB performs legacy handover execution. Otherwise, it triggers a NV-HO towards the MC-Home.

The key advantages and weak aspects of both algorithms can be summarized as follows. The NV-RSRQ algorithm is expected to achieve significant performance improvements at the UEs, since it enables the UEs to connect to the base station providing
the most favorable signal quality conditions (regardless the operator to which it belongs to). At the same time, this can be the cause of increased leasing costs for the operator A as well as excessive traffic offloading towards the network of other operators. On the contrary, the NV-Offloading algorithm is expected to compensate these weaknesses as it performs a NV-HO only if the serving eNB is overloaded. Such an approach is expected to lower the number of NV-HOs in a network-wide scale and reduce the leasing costs of operator A.

3.2.2 NV Decision

The NV decision phase is performed in step 5 (Figure 12). In this phase, the MC-Home decides on the most appropriate attachment point for the tagged UE in the cellular network of other operators (NV-HO). This phase is performed only if the home operator has established SLA agreements with other cellular operators.

To complete this phase, the serving eNB should provide the MC-Home with all information required to take the NV decision. Such information may include the characteristics of the ongoing UE connections, the list of reported measurements that triggered the NV-HO procedure, the load status at the serving eNB, the identity of the UE, and so on. In the sequel, it is assumed that the NV decision is taken based on the RSRQ status of the OpeNBs that are in proximity of the UE. Accordingly, upon NV-triggering, the MC-Home forwards a NV-HO request towards the OpeNB that belongs to the list of measurements provided by the serving eNB and attains the highest RSRQ for the tagged UE.

3.2.3 SLA Monitor

In this phase, the MC-Host checks if the NV-Request of the home operator is in line with the established SLA agreements. The function of this phase can be easily understood with by considering the following exemplary use case. It is assumed that a third part infrastructure provider, i.e. the operator B, lets multiple network operators to lease its network infrastructure for the duration of a specific event, i.e. a sport event or a concert, in a specific geographic area. Under this model, one operator can improve its network capacity by having access to the additional resources from the base stations belonging to Operator B. In this scenario, the SLA Monitor should verify that the shared resources are allocated to each operator in a fair manner, without exceeding the limits specified in the SLA agreements. If the SLA monitor phase is successful, the MC-Host forwards the NV HO request of the home operator (operator A) to the target OpeNB.
3.2.4 NV Admission Control

In this phase, the target OpeNB decides on whether to accept the NV-HO request, or not. The role of the NV Admission Control is to estimate the resource availability at the target OpeNB and infer on whether the target OpeNB can satisfy the quality of service (QoS) required by the tagged UE. In this phase, the OpeNB also estimates the impact of admitting a UE from another operator on the performance of the UEs that belong to host operator (Operator B). Note that the NV Admission Control phase can either give priority to the UEs of the host operator, or treat all attached UEs with the same priority, or perform cell-load based admission control.

3.3 Simulation Analysis

In this Section, system-level simulation results are presented to assess the performance gains attained by the employment of the proposed NV-based multi-tenant architecture in the LTE/LTE-A network. The considered scenario consists of a geographical area that includes two overlapped LTE-A networks owned by two different operators: Operator A and Operator B. The network of Operator A consists of 7 macro cells with inter-site distance of 1 km, while the network of Operator B consists of 7 macro cells that overlap the coverage of the macrocell network of Operator A.

Moreover, it is considered that Operator B owns a number of femtocells that are uniformly distributed within the coverage of the macrocells owned by Operator B. In more detail, the 3GPP 5X5 grid model [48] is used for defining the deployment of femtocells inside the buildings of a suburban area. It is also assumed that each building contains five femtocells that are uniformly distributed within the apartments of each building. The number of buildings is adapted depending on the desired femtocell density in the network and is used as the x-axis parameter in the simulation campaigns.

In the sequel, it is assumed that the Operator A acts as the home operator, whereas the Operator B acts as the host operator. Moreover all base stations of Operator B, including both macrocells and femtocells, are assumed to support the OpeNB functionality and operate in the same frequency. It is also assume that a fixed number of pedestrian UEs are uniformly distributed within the coverage area of each eNB operated by operator A (40 UEs/ eNB). A similar assumption is made for the UEs registered to operator B (30 UEs/ Macro Cell). The presented results are derived assuming that each UE sustains an H.264 video flow encoded...
at 440 kbps with a maximum delay constraint of 100ms [48]. The simulations results were derived by using the LTE-Sim simulator [48]. The rest of the simulation parameters are summarized in Table II.

### Table II: Simulation parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deployment</strong></td>
<td></td>
</tr>
<tr>
<td>#eNBs (Op.A)</td>
<td>7</td>
</tr>
<tr>
<td>Inter-site distance (Op.A)</td>
<td>1 Km</td>
</tr>
<tr>
<td>eNB bandwidth</td>
<td>10 MHz</td>
</tr>
<tr>
<td>#macro OpeNBs (Op.B)</td>
<td>7</td>
</tr>
<tr>
<td>Inter-site distance macro OpeNBs (Op.B)</td>
<td>1 Km</td>
</tr>
<tr>
<td>UEs density Operator A</td>
<td>40 UEs/Macro Cell</td>
</tr>
<tr>
<td>UEs density Operator B</td>
<td>30 UEs/Macro Cell</td>
</tr>
<tr>
<td>#Building per macro OpeNB</td>
<td>0 to 40</td>
</tr>
<tr>
<td>Building Layout</td>
<td>3GPP 5X5 model [15]</td>
</tr>
<tr>
<td>#Femtocells per Building</td>
<td>5, uniformly distributed</td>
</tr>
<tr>
<td>#Femtocells in the system</td>
<td>0 to 1400</td>
</tr>
<tr>
<td>OpeNBs bandwidth (Op.B)</td>
<td>10 MHz, (macro and femto)</td>
</tr>
<tr>
<td>Mobility Model</td>
<td>Random Way Point Model [15]</td>
</tr>
<tr>
<td>UE Speed</td>
<td>Pedestrian (3 km/h)</td>
</tr>
<tr>
<td><strong>System Model</strong></td>
<td></td>
</tr>
<tr>
<td>PATH Loss Model</td>
<td>$L = 128.1 + 37.6 \log_{10}(R)$ [15]</td>
</tr>
<tr>
<td>Fading Model</td>
<td>Jakes model [15]</td>
</tr>
<tr>
<td>Scheduler Downlink</td>
<td>MLWDF [15]</td>
</tr>
<tr>
<td>Application Traffic Model</td>
<td>Video Streaming (Download) encoded at 440 kbps</td>
</tr>
</tbody>
</table>

In the sequel, the performance of a legacy LTE network are compared with that of an LTE network that employs the proposed NV-based multi-tenant architecture. For brevity, the first scenario is referred to as the baseline scenario, i.e. the one where the proposed architecture is not employed. On the other hand, depending on the NV triggering and admission control algorithms adopted, the performance of the proposed NV-based architecture are examined under three different NV scenarios.

In the first scenario, tagged as as the “NV-RSRQ scenario”, the NV-RSRQ algorithm is employed as described in the Section 3.2 for the NV triggering phase, and it is considered
that the OpeNBs accept all the HO requests from Operator A in the NV Admission Control phase. In this scenario, all UEs are admitted as long as they the target OpeNB has enough resources to satisfy their QoS requirements. In the second NV scenario, tagged as as the “NV-Offload” scenario, the NV-Offloading algorithm is employed as described in Section 3.2.1 for the NV triggering phase and, once again, it is assumed that the target OpeNBs accept all NV-HO requests from Operator A (as long as they can support the QoS requirements). In the third NV scenario, tagged as the “NV-AC” scenario, the NV-Offloading algorithm is employed as described in Section 3.2.1 for the NV triggering phase, and it is considered that the NV-HO requests from Operator A are accepted only if the traffic load at the OpeNBs is lower than a specific threshold, i.e. the 85% of the total PRB utilization.

In the sequel, different plots are presented for assessing the performance of the UEs that are registered in the home operator (Figure 13) and the performance of the UEs that are registered in the host operator (Figure 14). Nevertheless, in the first plot (Figure 13) the performance of the UEs that are registered in the home operator and continue to receive service from their home operator are distinguished from the performances of the UEs that are registered in the home operator and exploit the proposed NV-based architecture for receiving service from the host operator (Operator B). For brevity, the suffix ‘… : eNB users’ is used to refer to the first set of UEs and the suffix ‘… : OpeNB users’ is used to refer to the second set of UEs.

Figure 13 shows the average goodput of the UEs that are registered to the home operator, i.e. the throughput as measured at the application layer, for all scenarios under scope and for increasing number of femtocells in the host operator. As shown in Figure 13, the UE goodput performance for all scenarios that employ the proposed NV-based architecture, i.e. the ones with the prefix ‘NV-…’, outperform the UE goodput performance in the baseline scenario. This trend directly follows from the fact that the employment of the proposed NV-based architecture enables the cellular UEs to connect to cellular BSs with more favorable channel conditions, e.g. higher RSRQ, or the lower path loss. Notably, we observe that the employment of the proposed NV-based architecture, not only improves the goodput performance of the UEs that exploit the proposed NV-based solution, i.e. the ones offloaded to Operator B, but also improves the performance of the UEs that continue to receive service from Operator A. In fact, the performance of the UEs that continue to receive service from Operator A is improved at a higher rate compared to the one of the UEs that are offloaded to Operator B. This follows from the fact that the proposed NV-based solution
enables traffic offloading towards the host operator, leaving more resources for the UEs that continue to receive service from Operator A.

Figure 13: Average DL Goodput for UEs registered to Operator A

Referring to the performance comparison between the different NV-scenarios under scope, it is observed that the highest performance gains are achieved in the NV-RSRQ scenario for both types of UEs registered to Operator A. This behavior follows from the fact that the NV-RSRQ triggering algorithm favors the execution of handovers towards the base stations with more favorable channel conditions. The performance of the NV-Offload and the NV-AC scenarios are analysed in the sequel. Note that in both of these scenarios more restrictive policies are employed during the NV Triggering and NV Admission Control phases. Firstly, it is observed that the goodput performance of the UEs offloaded to Operator B is roughly the same for both the NV-Offload and the NV-AC scenario. On the contrary, a slightly better performance is observed for the UEs that continue to receive service from Operator B in the NV-Offload, as compared to the one of the NV-AC scenario. Interestingly, under low femtocell densities (left side of the plot), the performance of the UEs offloaded to Operator B, under both the NV-Offload and the NV-AC scenarios, is higher compared to that of the UEs that continue to receive service from Operator A. However, this behavior alters in higher femtocell densities, where an increased number of UEs from Operator A can be offloaded to the femtocells of Operator B, leaving more resources for the UEs that continue to receive service from Operator A.
Figure 14: Average DL Goodput for UEs registered to Operator B

Figure 14 examines the average goodput at the UEs that are registered in Operator B. As expected, the employment of the proposed NV-based architecture reduces the average goodput at the UEs registered in Operator B, as a result of the increased demand of network resources, i.e. increased number of served UEs in Operator B. This observation readily follows by comparing the performance of the baseline scenario to that of the NV-based scenarios. Figure 14 also reveals that the highest gains at the UEs of Operator A (NV-RSRQ in Figure 13) are attained at the cost of the highest performance losses at the UEs of Operator B (NV-RSRQ in Figure 14). Nevertheless, this mainly follows from the fact that the NV-RSRQ scenario assumes that the OpeNBs do not perform admission control at the Operator B. To this end, as shown in Figure 14, the employment of load-balancing based criteria during the NV triggering (NV-Offloading) and the NV admission control phases (NV-AC) can result in notable performance gains for the UEs registered to Operator A (Figure 13) without significantly deteriorating the performance of the UEs registered to Operator B (Figure 14).

Figure 15 shows the average downlink signal to interference plus noise ratio (DL SINR) for the UEs registered to the home operator (averaged over all users). As the number of femtocells increases, an enhanced DL SINR is experienced at the UEs registered to Operator A, including both the ones that continue to receive service from Operator B and the ones that are offloaded to Operator B. This performance improvement follows from the flexibility offered to the UEs registered to Operator A, to associate with base stations owned by different mobile network operators. As in Figure 13, the highest performance gains are
shown to be attained for the NV-RSRQ scenario (close to 1.5 dB), whereas notable performance gains are observed for the NV-Offloading and NV-AC scenarios as well (close to 1.2 and 1 dBs, respectively). At this point, it is important to note that in these simulations, it is assumed that the macrocell and the femtocell stations belonging to Operator B, operate in the same frequency. Hence, even higher performance gains would be expected if the femtocell and macrocell base stations in Operator B utilize different frequencies.

![Average DL SINR for UEs registered to Operator A](image)

**Figure 15: Average DL SINR for UEs registered to Operator A**

Figure 16 shows the average DL SINR at the UEs registered to the host operator, for the different scenarios under scope and for increasing femtocell density per macrocell. As expected, the admission of additional UEs in Operator B, i.e. the ones utilizing the proposed OpeNB operation, reduces the average DL SINR for all UEs registered to Operator B. Nevertheless, as compared to the performance gains attained for the UEs registered at the home operator (Operator A), the performance loss is comparably lower, and strongly depends on the NV-scenario under scope. In more detail, the NV-RSRQ scenario is shown to reduce the average DL SINR at the UE registered in Operator B by up to 0.5 dB. On the contrary, both the NV-Offloading and the NV-AC scenarios are shown to attain slightly lower performance as compared to the baseline scenario, where the NV-AC can be said to attain roughly the same performance for all densities under scope.
Figure 16: Average DL SINR for UEs registered to Operator B

Figure 17: Average packet delay for UEs registered to Operator A

Figure 17 depicts the average end-to-end delay experienced by the UEs registered to the home operator (Operator A), as measured at the application layer. Firstly, it is observed that the employment of the proposed NV-based architecture significantly reduces the end-to-end application-layer delay at the UEs registered to Operator A for all NV scenarios under scope. Secondly, it is observed that the performance gains are proportional to the number of femtocells available from the host operator (Operator B). Thirdly, Figure 17 reveals that the end-to-end application-layer delay at the UEs is roughly the same for all the NV scenarios under scope when the femtocell density at the host operator is medium to high, i.e. higher than 160 femtocells per macro OpeNB.
Another interesting observation is that apart from increased goodput at the application layer (Figure 15), the employment of the proposed NV-based architecture results in substantial reduction of the end-to-end application-layer delay at the UEs (Figure 17). This reduction reaches up to 60% as compared to the baseline scenario. Such performance gains can significantly enhance the experience of the end-user, upon reception of delay-sensitive services, fully capitalizing the performance gains following from shifting NV to the access network of the LTE-A system. It should be noted that the performance improvement, in terms of end-to-end application-layer delay, not only follow from the enhanced goodput attained at the UEs, but also from the flexibility that enables the UEs to associate with the closest base station in proximity (even if it belongs to a different operator). Besides, in lower loads, the packet scheduler at the BSs of the home operator can better handle the packet flows of the UEs that remain in the home operator. This effect further reduces the queue processing time, decreasing the overall packet delivery latency as well.

![Figure 18: Average packet delay: UEs registered to Operator B](image)

The impact of offloading a certain number of UEs to the host operator, in terms of end-to-end application layer delay at the UEs registered to the host operator is examined in Figure 18. As expected, the end-to-end application-layer delay at the UEs registered to Operator B is reduced for higher femtocell deployment densities (baseline scenario). This behavior is common for all scenarios under scope, since the deployment of additional femtocell stations at the host operator, increases the network density and reduces the mean distance to the nearest base stations for all UEs served by the network of Operator B. Interestingly, even though the employment of the proposed NV-base architecture is shown to
increase the average end-to-end application-layer delay for the UEs registered to Operator B, as the femtocell deployment density increases, this performance deterioration is comparably smaller than the performance gains attained at the UEs registered to Operator A (Figure 17). The employment of the NV-RSRQ scenario, which has been shown to provide up to 21ms reduced delay for the UEs registered to Operator A, is shown to increase the end-to-end delay at the UEs registered to Operator B by up to 9ms. On the other hand, the performance of the NV-AC and the NV-Offloading scenarios is shown to be roughly the same as compared to the baseline scenario (up to 2ms increase of the end-to-end delay). At this point, it is important to note that the performance deterioration at the UEs of Operator B will leave their QoS performance unaffected, since the maximum delay requirement for the assumed traffic type, i.e. delay-demanding video streaming, is 100ms.

![HO signaling rate for the UEs registered in Operator A](image)

**Figure 19: HO signaling rate for the UEs registered in Operator A**

Figure 19 shows the number of HO signals per second, a.k.a. HO signaling rate [49], that are exchanged for the UEs registered to Operator A, for all scenarios under scope and for increasing number of femtocells per macro OpeNB. Noticeably, when the mobile cellular network of Operator B is composed by macro OpeNBs only, i.e. when the number of femtocells per Macro OpeNB is equal to zero, the employment of the proposed NV-based architecture results in a small increase in the HO signaling rate for the UEs registered in Operator A (left side of the plot). This result indicates that the employment of the proposed NV-based architecture roughly affects the HO signaling performance of the two networks in the absence of femtocells at the host operator. On the other hand, as the number of
femtocells increases (in the host operator), the HO signaling rate (for the UEs registered in Operator A) also increases due to the presence of additional OpeNBs at the host operator. Note that the employment of the proposed NV-based architecture increases the HO signaling rate as well, due to the signaling procedure required for establishing the NV-based link (Figure 12). Nevertheless, as discussed above, the HO signaling overhead in Figure 19 is mainly increased due to the flexibility offered by the proposed NV-based architecture to utilize the femtocell infrastructure of the host operator (Operator B).

It should be noted that the VeNB instantiation and tunnel establishment phases are performed only the first time that a UE visits a tagged OpeNB of the host operator. This observation plays a key role in measuring the performance of the NV-based scenarios in Figure 19, as it significantly reduces the number of steps required for performing a NV-based HO.

### 3.4 CONCLUSIONS

A non-invasive NV-based architecture for LTE/LTE-A networks that enables mobile operators to shift the SDN/NFV technology to their access network has been proposed. Extensive system-level simulations have been used to assess the performance of the proposed NV-based architecture and compare it to a legacy LTE/LTE-A network. In more detail, the simulation results have shown that, even under the employment of simple NV-triggering algorithms, the proposed architecture can significantly improve the goodput and end-to-end delay performance at the UEs registered to the home operator. Nevertheless, these performance gains are attained at the cost of a slight increase in HO signaling rate between the two operators and small performance deterioration at the UEs registered to the host operator. Notably, simulation results have also shown that the performance of the UEs registered to the host operator can remain roughly unaffected if simple admission control schemes are employed at the base stations that are endowed with multi-tenant capabilities. The future research work aims to propose more effective resource-sharing strategies among the operators involved in the proposed NV-based architecture and shed more light on how the proposed NV-based solution can be exploited in the context of evolving the LTE-A network architecture towards the 5G RAN architecture.
4. Programmable RAN (NEC)

4.1 Concept Description/Programmable RAN for TD-LTE

The fast adoption of smartphones and tablets combined with the widespread deployment of high speed mobile broadband networks, have led to an evolution of diverse mobile services, creating a tsunami of data traffic [51]. Mobile applications such as social media and cloud services have changed the way humans communicate and acquire information from the Internet, being also more interactive due to “always on” features, with higher uplink demands [52]. Equivalently, Machine Type Communication (MTC) applications need to upload or exchange high data volumes raising certain Quality of Service (QoS) demands, besides the yet more diverse uplink traffic patterns. Hence, emerging mobile applications and services require an increased degree of network resource elasticity to assure effectively different QoS demands.

Time-Division Long Term Evolution (TD-LTE) supports an unpaired frequency band, wherein uplink (UL) and downlink (DL) are separated in the time domain. Each TDD frame may support a potential variety of UL/DL ratio, following 7 different configurations as specified in [53], offering resource diversity for emerging data applications. Despite such UL/DL resource flexibility, the initial deployment of TD-LTE involves a synchronous frame configuration across certain network regions with all the cells offering an identical UL/DL ratio in order to avoid cross-slot interference [54]. Such network arrangements effectively limit the advantage of UL/DL resource flexibility since it imposes restrictions among neighboring cells. In response, the 3rd Generation Partnership Project (3GPP) focused on a set of enhancements to address interference and accommodate traffic adaptation [55]. These efforts introduce cell specific mechanisms in the sense that UL/DL configuration change takes place in each cell individually in a distributed manner with the goal to optimize uplink and downlink network resources to best serve the local traffic demands, assuming that user equipment (UE) are associated with a single cell at a given time.
Table 3: TD-LTE Uplink-downlink Frame Configurations

<table>
<thead>
<tr>
<th>UL/DL Configuration</th>
<th>Sub-frame Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>DL</td>
</tr>
<tr>
<td>1</td>
<td>DL</td>
</tr>
<tr>
<td>2</td>
<td>DL</td>
</tr>
<tr>
<td>3</td>
<td>DL</td>
</tr>
<tr>
<td>4</td>
<td>DL</td>
</tr>
<tr>
<td>5</td>
<td>DL</td>
</tr>
<tr>
<td>6</td>
<td>DL</td>
</tr>
</tbody>
</table>

However, neighboring cells may introduce overlapping regions, where UEs can potentially utilize resources, i.e. TDD sub-frames, from different cells provided that interference is considered. When neighboring cells adopt a distinct UL/DL configuration, such overlapping regions may offer yet another UL/DL ratio forming a customized unique type of frames derived from the UL/DL configurations of the involved cells, introducing in this way a notion of cell that is virtual, i.e. with no physical infrastructure. The benefits of virtual cells in enhancing the resource allocation efficiency of TD-LTE were shown in [56], which enables mobile operators to match closer the offered resources with the service demand. Nevertheless, such analysis focused on demonstrating the performance potential of a virtual cell considering a simple scenario of two base stations or 40 evolved Node Bs (eNBs) in 3GPP terminology.

SDN based network management in 3GPP networks has received significant attention in the research community indicating potential to provide enhanced flexibility and efficient network resource management [57]. This paper introduces the mechanisms and network management architecture that can facilitate a broader adoption of virtual cells in TD-LTE networks. It encounters also a closer coordination with cell specific adaptive UL/DL frame re-configuration, i.e. enforcing a selected frame re-configuration on particular cells that may enable an improved virtual cell formation, without compromising the performance of other existing users. Considering the network management architecture, this paper adopts the SDN paradigm to facilitate a fine-grained network resource control based-on a global network view, which also enables application and service providers to acquire QoS and resources via the Application-Controller Plane Interfaces (A-CPI) [58]. Although this later feature is not explicitly explored into this work, the proposed mechanisms can accommodate
on-demand resource allocation, which reflects the generic case for OTT resource acquisition via the use of the A-CPI.

4.2 Related Work

In the initial TD-LTE deployments, each cell supported an identical TDD frame that matches best the estimated long term UL/DL traffic demands. This synchronized operation across a network region of eNBs offers a uniform constant UL/DL ratio providing a simple solution for avoiding cross-slot interference. Efforts towards improving the flexibility of such a scheme concentrate on handling interference, which is the fundamental constraint against a dynamic form of TD-LTE in where each cell is capable to adopt a different TDD frame and change it over time based-on evolving traffic demands. In [54], [59], the challenges and benefits for adopting a flexible TD-LTE scheme are analyzed focusing on mechanisms that dynamically change the UL/DL configuration on each cell at different timescales to match the UL/DL traffic demands accommodating power control and/or clustering mechanisms for avoiding interference. Our proposal adopts a similar dynamic TD-LTE scheme, employing power control mechanisms to accommodate interference mitigation.

Recent advancements on TD-LTE investigate mechanisms that aim to optimize the UL/DL ratio selection for each eNB, considering a variety of different constraints. An asymmetric assignment of UL and DL is investigated in [60], which aim to reshape interference channels exploiting efficiently the available backhaul resources. In [61] a cooperative decentralized mechanism is introduced that provides a local optimal solution considering instantaneous data rate conditions and traffic demands, which are exchanged via reliable low-rate signaling among neighbor cells.

An alternative approach that aims to minimize the information exchange among neighboring cells by enabling eNBs to perform autonomously an UL/DL ratio optimization based-on a game theoretic method is introduced in [62], considering UL/DL delays in relation with traffic load, interference and flow-level dynamics. Optimizing further the selection of the TDD frame considering the evolving Quality of Service (QoS) demands in terms of bit 95 rate guarantees and packet delay is elaborated in [63]. The potential gains in throughput and reduced packet delays once each TDD slot is freely assigning as UL or DL instead of using one of the 7 pre-determined TDD frames is explored in [64], assuming the use of Interference Rejection Combining (IRC) capable receivers to handle cross-slot interference.
However, such proposals consider individual eNBs focusing on the means of providing a dynamic UL/DL ratio in an autonomous way considering a set of various optimization parameters. Unlike these cell centric approaches, our proposal stretches beyond a single cell, i.e. local optimization, exploiting efficiently the resource diversity within overlapping cell areas as elaborated in [56]. This enables UEs to utilize sub-frames from multiple eNBs, i.e. forming virtual cells that offer customized TDD frames, which match best their UL/DL traffic demands. Performing radio resource management that combines dynamic TDD re-configuration with virtual cells in a distributed manner is suboptimal. Hence, centralized intelligence is recommended for efficient interference mitigation considering traffic dynamics. A similar approach referred to as V-Cell, offers a combination of heterogeneous radio resources, i.e. macro-cells, pico and femto cells, as a resource pool to UEs, which perceive such an access as a logical single macro-cell [65]. The network resource management is performed by an SDN controller, which maintains a logical global view of the underlying network in order to efficiently schedule resources across the entire pool of physical radio elements.

An equivalent SDN-based RAN management architecture that relies on abstracting the RAN resources and using them as a single virtual wireless access is also analyzed in [66], which also provides more details about the SDN controller including the main associated functions and the control plane mechanisms. Effectively such components enable RAN programmability as considered in [67], which analyzes the SDN impact of separating the control and data planes in easing the management of heterogeneous networks considering mobility and QoS-aware network operation. In fact, the SDN architecture offers the N-API towards application providers facilitating on-demand QoS provision. In this paper a similar SDN paradigm is employed offering RAN programmability, which is realized by enabling a unified control for provisioning a dynamic TDD frame configuration at selected eNBs with the potential of forming virtual cells based-on the user resource demand. Whilst our focus is on network programmability based-on network measurements and resource utilization, the proposed SDN architecture may potentially offer a customized UL/DL ratio upon a request via the N-API, allowing applications to program the network and provision resources for particular services within certain times. Such a feature may advance the network customization offering flexibility, scalability and easy deployment of new services and enhanced performance of TD-LTE systems.

This section describes the main components for realizing the proposed elastic TD-LTE resource management mechanisms. Initially, the SDN architecture that manages the TD-LTE network is described, followed by the resource management logic for creating virtual cells and the TD-LTE resource allocation algorithm for re-programming the network configuration.

4.4 Virtual Cell Concept

The concept of virtual cell allows users to utilize subframes from multiple eNBs, enabling the formation of a customized virtual frame by deriving specific subframes from different eNBs that can react best the user UL and DL transmission demands [56]. The flexibility offered by this feature allows the resolution of pseudo congestion, while enhancing the user performance. Such a technique could also be beneficial for the users that are outside the virtual cell region as it could free up additional resources for them to achieve their desired QoS. In addition to this this technique not only exploits the spatial domain of conventional load balancing but also the time domain to dynamically configure the cell setup.

![Figure 20: A Simple Example of the Virtual Cell Concept](image)

The adjacent cooperating eNBs in virtual cell concept appears to be as one logical cell with each eNB offering a different UL/DL configuration. This provides the capability to support multiple and diverse applications within smaller geographical regions. It is worth noting that the UEs cannot utilize UL and DL sub-frames within the virtual cell region at the same time because of device and hardware restrictions. The process of utilizing sub-frames...
from different eNBs requires enhanced mechanisms to synchronize UEs and align their transmit/receive modes accordingly.

Therefore synchronization is required between the eNBs involved in the virtual cell formation to ensure that the data towards and from the UE appears as a single stream hence the virtual cell requires additional signaling mechanisms for control purposes. A simple example of the virtual cell concept is illustrated in Figure 20 where, a UE residing within the virtual cell region is utilizing UL resources from eNB A and DL resources from eNB B to match its traffic demand. Specifically, the congestion occurs as the UE has high UL and relatively low DL resource demand and hence cannot be served solely by any of the depicted eNBs without experiencing congestion. This pseudo congestion can be avoided by enabling UE to utilize resources from both the eNBs. Here, both the eNBs have resources available but in the opposite transmission directions [68], therefore, the UE utilizes resources from both the eNBs in the direction opposite to the direction of congestion.

4.5 SDN-based Network Management Architecture

The SDN-based system aims to perform network resource management and TD-LTE programming considering adjustments on selected TDD frames forming also virtual cells. The objective is to abstract the control plane from individual eNBs and logically centralize it, resulting in a collective resource management and control of eNBs’ resources, as described in [66].

To accommodate such a vision, the OAM accompanied by the Data-Controller Plane Interfaces (D-CPI), can facilitate periodic or on-demand RAN state updates in order to help the SDN controller to form a global network view [66]. In particular, the OAM can provide the SDN controller with RAN topology information, UL and DL load and Key Performance Indicators (KPIs), e.g. handover failures, latency, throughput, etc. as specified in [69]. The D-CPI may additionally provide the SDN controller with certain information related to specific rules including monitoring, such as interference levels and the TD-LTE frame configurations per eNB [70].
With a global visibility across the RAN, the SDN controller can enhance the resource allocation, enabling virtual cells by adjusting dynamically the power and subcarrier allocation profile of each eNB. The SDN controller can assess the impact of a TDD frame reconfiguration in the overall performance of the entire network and selectively enforce certain TDD frame changes at specific RAN locations, which otherwise won't be performed through the use of the local adaptive TD-LTE [54]. In addition, scalability is improved since TD-LTE re-configuration algorithms that may require a significant amount of data can be executed at the SDN controller rather than at eNBs, which have limited computational capacity, avoiding also extensive distributed signaling among multiple eNBs and backhaul elements, while assuring stability.

When the SDN controller is notified or it identifies the need for a change in the TD-LTE arrangement, it ties to determine if a TD-LTE frame re-configuration on particular eNBs and/or the use of virtual cells can enhance the resource utilization. To accomplish this, SDN controller is making use of the global network state provided by the RAN info-base and executes the provided algorithm and network orchestration policy. It should be noted that the algorithmic logic and orchestration policy can potentially be programmed via the N-API or Application-CPI (A-CPI), where operators have this degree of freedom and flexibility. Once, the SDN controller determines the new TD-LTE resource allocation solution it communicates the essential changes on the TD-LTE configuration back to the corresponding eNBs via the
C-DPIs. Effectively, this may alter the transmission power associated with particular TDD sub-frames to assure interference mitigation among neighboring eNBs and facilitate the creation of virtual cells. Whilst left for further study and only considered here for the architecture completeness, the SDN controller may also allow application provider to program the RAN via A-CPI, enabling QoS provision for particular services. An overview of the SDN architecture that elaborates the main elements of the elastic TD-LTE mechanism including their interaction is provided in Figure 21.

Since, network traffic interference conditions may fluctuate significantly even for short time periods, especially within the relatively small virtual cell regions, due to user mobility, additional distributed mechanisms should be considered. Such mechanisms may relax the workload on the SDN controller, allowing longer time scale tolerance on the TD-LTE configuration decisions provided to the RAN. Hence, local radio resource adjustments should complement the ones provided by the SDN controller as long as no neighbor eNBs are affected.

4.6 SINR Analysis and Effective Capacity Allocation for Virtual Cells

TD-LTE systems are particularly sensitive to interference, especially when neighbor eNBs follow a different UL/DL ratio, e.g. in the case of cell specific adaptive reconfiguration and virtual cell, due to cross-slot interference, caused by eNBs that directly interfere or among UEs in close proximity that communicate in the opposite transmission direction and may receive interference from each other [71]. Due to physical limitations of the radio frequency frontend at eNBs and UEs, there is the need to avoid such cross-slot interference. In our proposal we achieve this via the use of power control and ensure that the interference introduced cause negligible degradation in the user performance by computing the SINR considering the aforementioned interference phenomena.

The SINR is computed for each user associated with a particular eNB over the RBs assigned to it for transmitting the data. The expression for SINR for the UL and DL direction is given by:

\[
\gamma = \frac{P_{rx}}{I + N_0}
\]

(38)

Where, \( P_{rx} \) is the received power in UL or DL direction respectively, \( I \) is the interference power and \( N_0 \) is the noise power. In the UL direction, UEs that transmit data
towards their serving cell over the RBs assigned by the MAC scheduler may possibly employ the same subset of RBs that are utilized in the neighbor cell at the same time. Therefore, the interference experienced by the received signal at the serving cell over the RBs should be considered while computing the SINR expression in the UL direction. In the DL, all the transmissions on the same subset of RBs coming from other eNBs are considered while computing SINR expression.

The $P_{rx}$ in UL or DL is given by:

$$P_{rx} = P_t \cdot \tau \cdot \left( \frac{d_0}{d} \right)^\phi \cdot \psi$$  \hspace{1cm} (39)$$

where, $P_t$ is the transmit power in UL or DL, is a unit less constant, which depends on the antenna characteristics and average channel attenuation, $d$ is the distance between the transmitter and receiver and $d_0$ is a reference distance for the antenna far field [71],[72]. $\phi$ is a pathloss exponent and $\psi$ is a Gauss-distributed random variable representing the shadowing effects in propagation with mean zero and variance $\sigma^2_\psi$. The interference power is given by:

$$I = \sum_{i \neq o} P_{t,i} \cdot \tau_i \cdot \left( \frac{d_0}{d_i} \right)^\phi_i \cdot \psi_i$$  \hspace{1cm} (40)$$

While deriving the SINR expression for the cell specific adaptive reconfiguration and virtual cell considering the UL and DL directions, the cross-slot interference, that may arise, should also be taken into account. The SINR expression for both cases in the UL direction for a user $i$ connected to a cell $l$ with respect to a set of neighboring cells $J_l$ can be expressed as:

$$\gamma_{i,l}(UL) = \frac{P_{UL}}{\sum_{k=1}^{J_l} I_{DL}(l,k) + \sum_{k=1,k \neq l}^{J_l} I_{UL}(l,k) + N_0}$$  \hspace{1cm} (41)$$

Where, $I_{DL}(l,k)$ is the interference power of the DL signal in the neighbor cell $k$, $k \in J_l$ observed at the serving cell $l$, $l \in L$. $I_{UL}(l,k)$ is the interference power of the UL signal in the neighbor cell $k$, $k \in J_l$ observed at the serving cell $l$, $l \in L$. $N_0$ is the noise power. The equivalent SINR expression in the DL direction is:

$$\gamma_{i,l}(DL) = \frac{P_{DL}}{\sum_{k=1,k \neq l}^{J_l} I_{DL}(l,k) + \sum_{k=1}^{J_l} I_{UL}(l,k) + N_0}$$  \hspace{1cm} (42)$$
Where, $I_{DL}(l,k)$ is the interference power of the DL signal from the neighbor cell $k, k \in J_l$ measured at the UE $i$ in the cell $l, l \in L$. $I_{UL}(l,k)$ is the interference power of the UL signal from an active UE operating in the neighbor cell $k, k \in J_l$ at the UE operating in the serving cell $l, l \in L$.

For enabling an efficient virtual cell formation, there is a need for a mechanism to ensure that the resource gain for a particular cell is not resulting in starving the users in another cell, i.e., there is no negative impact of virtual cell configuration on the neighbor eNB from which the resources are taken. We ensure this by performing capacity gain and loss calculations, considering the effective bandwidth model presented in [73], for calculating the throughput on the serving link as well as the potential virtual neighbor cell. The capacity of the link $S$ [b/s/Hz] in UL and DL is:

$$S_{i,k} = \min \left( B_{\text{eff}} \cdot \log_2 \left( 1 + \frac{\gamma_{i,k}}{\gamma_{\text{eff}}} \right), \ S_{\text{eff}} \right)$$

(43)

Where $B_{\text{eff}}$ is the bandwidth efficiency, $\gamma$ is the link-level SINR, and $\gamma_{\text{eff}}$ is the SINR efficiency, for a system with maximum spectral efficiency $S_{\text{eff}}$. For borrowing resources from neighbor eNBs and creating virtual cells, we consider a capacity gain and loss metric, similar to the one considered in [74]. Let $R_k$ be the total available resource blocks in eNB $k$, $\beta$ be the percentage of available resources that can be borrowed, and $\gamma_{i,k}$ be the SINR experienced by the user $i$ with eNB $k$. The capacity gain by borrowing resource and creating a virtual cell, $C_g$ [b/s] in UL/DL is given by:

$$C_g = R_k \cdot \beta \log_2 \left( 1 + \gamma_{i,k} \right)$$

(44)

The capacity loss $C_l$ [b/s] in UL/DL for the mean user $m$ of eNB $k$, having SINR $\gamma_{m,k}$ (UL/DL), due to the user $i$ borrowing the resources is given by:

$$C_l = R_k \cdot \beta \log_2 \left( 1 + \gamma_{m,k} \right)$$

(45)

The virtual cell is created with eNB $k$ for user $i$, if $C_g > C_l$. This condition ensures that the resource borrowing is done only when there are some capacity gains for the congested cell of UE$i$, and is limited by the factor $\delta \in [0,1]$. Since the capacity gain can never be higher than the loss in a macro cellular system with identical cell traffic load, and received signal...
strength based cell selection, we have used the limiting factor to dynamically control the virtual cell creation criteria. The mean user SINR $\gamma_{m,k}$ for eNB $k$ having $N_k$ UEs is given by:

$$\gamma_{m,k} = \frac{\sum_{u=1}^{N_k} \gamma_{u,k}}{N_k}$$  \hfill (46)

The virtual cell creation decision is taken per UE, depending on its link quality with its serving cell having congestion, as well as the strongest neighbor cell. For a user $i$, with eNB $o$ as its own cell (congested cell) having $R_{i,o}$ resource allocated to it and eNB $k$ as the virtual cell with $R_{i,k}$ resources borrowed from the cell, the total link level capacity of the user, $C_i$ [b/s] in UL/DL is given by:

$$C_i = \sum_{c=o,k} R_{i,c} \cdot S_{i,c}$$  \hfill (47)

If the capacity gain and loss condition is satisfied only then the required resources($\beta$), to resolve the congestion are borrowed otherwise the users are served via the standard cell specific adaptive reconfiguration process.

4.7 Elastic Resource Management Algorithm for TD-LTE

The resource management algorithm aims to perform elastic capacity allocation in UL or DL direction regulating the formation of virtual cells in order to enhance the users' performance and resolve potential pseudo-congestion problems. The algorithm also intends to maximize the throughput of low SINR users without causing negative effects on the QoS of other users, residing within the region of the serving cell or the surrounding neighboring cells.

The algorithm is executed at the SDN controller taking as input the global view of the network and the set of congested cells, i.e. cells that experience congestion for a time duration $t > T_{cong}$, where, $T_{cong}$ is the time period for congestion before an action is taken. The notion of congestion for the best effort traffic, where there is no strict requirement to maintain a specific data rate, is accounted considering the throughput. The throughput thresholds in UL and DL directions are used to detect congestion and trigger the algorithm in order to assess the situation and resolve the situation. The throughput thresholds in UL and
DL are selected considering the service delay, which should not exceed 300 ms according to [75] for the best effort traffic.

**Table 4: List of Variables used in the pseudo-algorithms**

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l \in L )</td>
<td>Cell belonging to a set of total network cells ( L )</td>
</tr>
<tr>
<td>( o \in O )</td>
<td>Cell belonging to a set of congested cells ( O )</td>
</tr>
<tr>
<td>( l[a] )</td>
<td>UL or DL average active user throughput in cell ( a )</td>
</tr>
<tr>
<td>( t_p(i) )</td>
<td>UL or DL active user ( i ) throughput</td>
</tr>
<tr>
<td>( t_{ph} )</td>
<td>UL or DL throughput threshold</td>
</tr>
<tr>
<td>( T_{cong} )</td>
<td>Time duration beyond which a cell with continuously limited resource is declared as congested</td>
</tr>
<tr>
<td>( k \in J_o )</td>
<td>Cell belonging to a set of neighboring cells ( J_o ) of a congested cell ( o )</td>
</tr>
<tr>
<td>( x_{J_o} )</td>
<td>Cell belonging to ( J_o ) with maximum ( R_k ) in UL or DL direction</td>
</tr>
<tr>
<td>( R_k )</td>
<td>Total available resources in a cell ( k )</td>
</tr>
<tr>
<td>( R_{req} )</td>
<td>Amount of resources needed to resolve congestion in ( o )</td>
</tr>
<tr>
<td>( \beta_l )</td>
<td>Available UL or DL resources of cell ( l ) that can potentially be used to form a virtual cell</td>
</tr>
<tr>
<td>( F[l] )</td>
<td>Current TD-LTE frame of cell ( l \in L )</td>
</tr>
<tr>
<td>( S_F )</td>
<td>Set of sub-frames that belong to a TD-LTE frame ( F[l] )</td>
</tr>
<tr>
<td>( F_n[l] )</td>
<td>New TD-LTE frame for cell ( l \in L )</td>
</tr>
<tr>
<td>( i \in U_o )</td>
<td>User belonging to a sorted set of active users ( U_o ) in an incremental ( \gamma_{l,n} ) order, residing in a congested cell ( o )</td>
</tr>
<tr>
<td>( vCell(l,F) )</td>
<td>Set of UEs assigned to the virtual cell region</td>
</tr>
<tr>
<td>( r_{x_{J_o}}(i) )</td>
<td>Resources allocated to user ( i ) once associated with a virtual cell borrowed from the selected neighbor cell ( x_{J_o} )</td>
</tr>
<tr>
<td>( \gamma_{l,n} )</td>
<td>SINR experienced by an active user ( i ) residing in a congested cell ( o )</td>
</tr>
<tr>
<td>( \gamma_{l,k} )</td>
<td>SINR experienced by an active user ( i ) from a neighbor cell ( k )</td>
</tr>
<tr>
<td>( \gamma_{m,k} )</td>
<td>SINR experienced by the mean active user of a neighbor cell ( k )</td>
</tr>
<tr>
<td>( C_g )</td>
<td>Capacity gain</td>
</tr>
<tr>
<td>( C_l )</td>
<td>Capacity loss</td>
</tr>
</tbody>
</table>

The algorithm initially examines the set of congested cells with the objective of forming virtual cells, selecting the optimal neighboring cell, which can offer the desired amount of resources. During this process the algorithm may also enforce a TD-LTE frame re-configuration to resolve potential pseudo-congestion, if that allows adequate resources for forming virtual cells. Once the cells that comprise the virtual cell region are selected, the
algorithm examines which users should be associated with such a virtual cells region considering the capacity gain and capacity loss. The variables used throughout the proposed algorithm are summarized in Table 4.

The pseudo-code of the algorithm that concentrates on the cell selection to form virtual cells is illustrated in Algorithm 1. In line 1 the algorithm collects the set of congested cells $O$. For each congested cell $o \in O$ it identifies its neighbor list $J_o$, from which it selects the neighbor cell referred to as $x_{(J_o)}$ with the maximum resource availability $R_k$ towards the congestion direction, which may either be on the UL or DL as shown in lines 3 and 4 respectively. Here, the goal is to identify a neighboring cell that can accommodate adequate resources in the desired transmission direction allowing the creation of a virtual cell, which fulfills both UL and DL demands. In this way, the algorithm tries to resolve congestion, while making the best use of the current network formation performing no changes to the TD-LTE network configuration.

If the selected cell $x_{(J_o)}$ is able to offer adequate potential resources $\beta_{x(J_o)}$ be used to form a virtual cell satisfying the resource request $R_{req}$ as shown in line 5, a virtual cell is
formed and the algorithm then selects the users to associate with such a region, allocating resources based on their location and interference levels, as elaborated in Algorithm 2.

Otherwise, the algorithm examines the entire per cell resources irrespective of the transmission direction $\sum_{\text{UL/DL}} R_k$, with the goal to identify a neighbor cell with the maximum total resource availability and stores its TD-LTE frame as $F[x_{(Jo)}]$ according to lines 8 and 10 respectively. It should be noted that such a neighboring cell even in cases where it cannot fully accommodate the resource request $R_{req}$, it can still provide the best solution toward resolving congestion, enhancing the average user performance depending on the location and interference conditions.

For such neighboring cell $x_{(Jo)}$, the algorithm in lines 11 to 19 tries to investigate whether enforcing a potential TD-LTE frame re-configuration may enhance the resource allocation towards the virtual cell, i.e. $\beta_{x_{(Jo)}}$, without compromising the average user performance, ensuring $x_{(Jo)}[t_p] > t_{Pth}$. In particular, in line 12 a TD-LTE frame re-configuration is performed, with the new TD-LTE frame $F_n[x_{(Jo)}]$ selected considering the current one, i.e. $F[x_{(Jo)}]$, with the minimum amount of sub-frames $S_F$ re-configured towards the congestion direction. For example if the current configuration $F[x_{(Jo)}]$ employs a DL/UL ratio of 8:1, i.e. configuration 5 in Table 3, and there is a need to enhance the potential of UL resources towards the virtual cell region, then re-configuring the minimum sub-frames towards the UL direction would result in a new TD-LTE frame $F_n[x_{(Jo)}]$ with an UL/DL ratio 7:2, i.e. configuration 4 in Table 3, while in the following iteration if the intention is to enhance UL resources even further would result in a $F_n[x_{(Jo)}]$ with an UL/DL ratio of 6:3, i.e. configuration 3 in Table 3.

A new frame is adopted by the system becoming the current one, continuing such an iterative process provided that the throughput change associated with the average user is still beyond the performance target threshold, i.e. $x_{(Jo)}[t_p] > t_{Pth}$, otherwise it ceases, breaking the iterative process, as shown in line 13 to 18. As stated before, once a neighbor cell is selected, potentially with a re-configured TD-LTE frame, a virtual cell is formed and then Algorithm 2 allocate resources towards specific users from the overloaded cell based on their location (i.e. the users residing within the virtual cell region) and interference conditions. The resource allocation is performed based on the user’s SINR levels and throughput using capacity gain and loss calculations.
In particular, Algorithm 2 initiates a set $vCell_{UE}$ to keep a record of the users assigned to the virtual cell region and creates a set of active users residing in a congested cell $U_o$, which is sorted in an incremental order according to the SINR experienced, $\gamma_{i,o}$, as shown in lines 1 and 2 respectively.

**Algorithm 2 Allocating Users to the Virtual Cell Region**

1: $vCell_{UE} \leftarrow \emptyset$
2: $U_o \leftarrow \text{sorted set of active users in an incremental } \gamma_{i,o} \text{ order}$
3: foreach $i \in U_o$ starting from $i$ with $\min(\gamma_{i,o})$
4: if $\text{tp}(i) < \text{tp}_{th}$
5: $C_g \leftarrow R_k \cdot \beta_{x(J_0)} \log_2(1 + \gamma_{i,k})$
6: $C_l \leftarrow R_k \cdot \beta_{x(J_0)} \log_2(1 + \gamma_{m,k})$
7: if $C_g > C_l$
8: //allocate user $i$ in the virtual cell region
9: $vCell_{UE} \leftarrow vCell_{UE} \cup i$
10: $\beta_{x(J_0)} \leftarrow \beta_{x(J_0)} - r_{x(J_0)}(i)$
11: else $C_g \leq C_l$
12: //do not interrupt user $i$
13: continue with next the user from $U_o$
14: end
15: end
16: if $\beta_{x(J_0)} \leq 0$
17: break foreach loop
18: end
19: end

For each active user starting from the one with the minimum $\gamma_{i,o}$, line 3, the algorithm checks if the user throughput $\text{tp}(i)$ is below or equal with the pre-determined threshold $\text{tp}_{th}$ in line 4. The rational for assessing first users with the minimum $\gamma_{i,o}$ is to try to improve the performance of users that are more in need since their SINR level is the lowest. For a user with throughput lower than the performance target, i.e. $\text{tp}(i) < \text{tp}_{th}$, the algorithm examines whether it is beneficial to allocate such a user to the virtual cell region, in lines 4 to 14. To accomplish this, initially the algorithm in lines 5 and 6 calculates the capacity gain $C_g$ and capacity loss $C_l$, as elaborated in section 4.3.3.

In case the capacity gain is greater than the capacity loss, the user is allocated in the virtual cell region. The algorithm adds that user to the $vCell_{UE}$ set and subtracts the allocated resources $r_{x(J_0)}(i)$ from the potential resources $\beta_{x(J_0)}$ that can be used within the virtual cell region. Otherwise, the user remains constant, i.e. uninterrupted, and the algorithm continues with the next user until all users $i \in U_o$ are considered or the $\beta_{x(J_0)}$ resources are exhausted. The algorithm returns the set of users $vCell_{UE}$ that should be allocated in specified virtual cell
regions, the neighboring cells involved and any enforced TD-LTE frame re-configuration associated with a particular neighbor cell. It is also worth noting that the users’ allocated resource from the virtual cell region could potentially free up resources for other users residing outside the virtual cell region and help towards resolve congestion in the problematic cell, thereby improving the overall network performance.

4.8 System Simulation and Result Analysis’

We carried out event based system level simulations in Matlab to evaluate the performance of the proposed SDN based elastic resource management solution that introduces virtual cells in TD-LTE network. We considered a standard 19-site and 3-sector hexagonal network layout, altogether forming 57 cells and adopted the evaluation methodology defined in [76]. In particular, UEs are randomly distributed in the service area and can access the system following a Poisson traffic model, with a mean arrival rate $\lambda$. Each UE accessing the system is capable of transmitting a file of size 0.5MB in UL and DL at different transmission time intervals (TTIs), assuming that traffic is generated randomly in uplink and downlink direction.

Table 5: System Simulation Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>eNB ISD</td>
<td>500m</td>
</tr>
<tr>
<td>System Bandwidth</td>
<td>10 MHz</td>
</tr>
<tr>
<td>Duplexing Scheme</td>
<td>TDD</td>
</tr>
<tr>
<td>eNB Max Tx Power</td>
<td>46 dBm</td>
</tr>
<tr>
<td>eNB Antenna Gain</td>
<td>15 dB</td>
</tr>
<tr>
<td>UE Total Tx Power</td>
<td>23 dBm</td>
</tr>
<tr>
<td>UE Antennal Gain</td>
<td>0 dB</td>
</tr>
<tr>
<td>Path loss Model</td>
<td>$128 + 37.6 \log_{10}(R)$, R in Km</td>
</tr>
<tr>
<td>Spectral Efficiency, Seff</td>
<td>4.0</td>
</tr>
<tr>
<td>Number of RBs, NRB</td>
<td>50</td>
</tr>
<tr>
<td>PRB size, RBs</td>
<td>180 kHz</td>
</tr>
<tr>
<td>Bandwidth efficiency, Beff</td>
<td>0.65</td>
</tr>
<tr>
<td>SINR efficiency, SINReff</td>
<td>0.95</td>
</tr>
<tr>
<td>File size (FS)</td>
<td>0.5 MB</td>
</tr>
<tr>
<td>$T_{RT_{thresh}}$(UL/DL)</td>
<td>$(0.5/1)$ Mbps</td>
</tr>
</tbody>
</table>

The system load is controlled by varying the user arrival rate that represents the average number of users accessing the system for transmitting and/or receiving a file. The traffic load of a cell is measured based on the number of active users at a given time and the
number of resources used. The detailed simulations parameters are summarized in Table 5 and are based-on the 3GPP LTE system specification [76].

We considered a scenario where at any given time; any random cell in the network may experience very high traffic in UL or DL leading to pseudo-congestion, while the other regions in the network carry relatively medium/low traffic. Considering this traffic scenario, we compared the following TD-LTE frame configuration schemes including:

- Static configuration 1, where all eNBs employ the same UL/DL configuration, with a subframe ratio of 60% DL and 40% UL.
- Cell specific adaptive reconfiguration, where the UL/DL subframe ratio is dynamically selected from the set of seven potential TD-LTE frame configurations. The selection of a suitable UL/DL configuration for individual cells is based on estimations of the uplink and downlink traffic demands as detailed in [54].
- Virtual cell, which utilizes resources from more than a single cell. Virtual cells are created by the SDN controller, which may optionally enforce an UL/DL reconfiguration to a particular cell in order to secure adequate resources for the virtual cell region. Under the virtual cell scheme eNBs can still perform locally a cell specific adaptive reconfiguration.

![Figure 22: Throughput Downlink bits/second](image)

The gains in the throughput performance can be observed in Figure 22 for DL and Figure 23 for the UL, which shows the throughput cumulative distribution function (CDF) graphs comparing the three aforementioned configuration schemes. From both figures it is
obvious that the Virtual Cell Configuration achieves significant gains in throughput compared to the Cell Specific Adaptive Reconfiguration and Static Config. 1, considering both users that have low SINR conditions that may reside at the cell edge and users with relatively higher SINR levels, that may reside in the inner cell and surrounding regions.

**Figure 23: Throughput Uplink bits/second**

These gains are clearly evident in Figure 24 and Figure 25, which shows the comparison between the throughput of low SINR users and the mean user throughput in the three schemes in DL and UL direction respectively.

**Figure 24: Mean and Low SINR Users’ Throughput Normalized to the Virtual Cell Configuration in Downlink**
The throughput of Cell Specific Adaptive Reconfiguration and Static Config. 1 is normalized to the Virtual Cell Configuration which helps in visualizing the gains observed by the virtual cell scheme in comparison with the Cell Specific Adaptive Configuration and Static Config. 1. Clearly, both schemes that provide a flexible UL/DL configuration, i.e. the Cell Specific Adaptive Reconfiguration and Virtual Cell Configuration, result in higher overall throughput, thereby outperforming the Static Config. 1, improving the performance of low SINR users.

Figure 25: Mean and Low SINR Users’ Throughput Normalized to the Virtual Cell Configuration in Uplink

We can observe from Figure 24 and Figure 25 that the Virtual Cell Configuration provides around 25% improvement in the low SINR user throughput in DL and around 20% in the UL compared to the Cell Specific Adaptive Reconfiguration, while 35% and 30% compared to Static Config. 1, in the DL and UL respectively. For the mean user throughput the Virtual Cell Configuration shows an improvement of 10% in the DL and around 6% in the UL compared to the Cell Specific Adaptive Reconfiguration, while around 16% and 12% compared to the Static Config. 1 in DL and UL respectively. The slight improvement in the mean user throughput demonstrates the fact that applying the virtual cell scheme has no negative impact on the overall performance of the users in the system.

Figure 26 and Figure 27 illustrate the delay CDFs of the aforementioned three schemes for the DL and UL respectively. It can be observed that the Virtual Cell Configuration achieves significant gains in transmission delay compared to the Cell Specific Adaptive Reconfiguration and Static Config. 1 both for low SINR users, which may reside at the
celledge and other users who experience relatively higher SINR levels that may reside in the inner cell and surrounding regions.

The gains of the transmission delay can also be visualized in Figure 28 and Figure 29 for the DL and UL direction respectively. As expected, the results are aligned with the throughput gains, as throughput and delay are closely related metrics. The delay measures of Cell Specific Adaptive Reconfiguration and Static Config. 1 is normalized to the delay of the Virtual Cell Configuration. From Figure 28 and Figure 29 it is observed that low SINR users in the Virtual Cell Configuration introduce around 25% and 20% less delay than the low
SINR users in cell Specific Adaptive Reconfiguration, while 35% and 30% compared to the Static Config. 1 in DL and UL respectively. It is also noted that the mean transmission delay for Virtual Cell Configuration is reduced 10% and 6% in the DL and UL direction compared to the Cell Specific Adaptive Reconfiguration and around 16% and 12% compared to the Static Config. 1.

In summary, it is evident from the performance analysis that the Virtual Cell Configuration outperforms the state of the art mechanisms mainly because it can dynamically allocate sub-frames within neighboring cells’ overlapping regions for the UL or DL directions according to the real time traffic demands. This provides enhanced flexibility by allowing the system to program the network resources on-demand considering a global network view addressing users' demands at particular cell areas and resolving pseudo congestion.

Figure 28: Mean and Low SINR Users’ Delay Normalized to the Virtual Cell Configuration in Downlink

Figure 29: Mean and Low SINR Users’ Delay Normalized to the Virtual Cell Configuration in Uplink
It is also worth noting, that the users residing in the virtual cell region and served via multiple eNBs could potentially free up resources for other users in the cell, which may experience enhanced SINR improving the overall network performance.

With the help of SDN based resource management and the use of virtual cells, operators can maintain a tight control over the network with the ability to flexibly allocate resources on-demand, not only to the end users but also to the OTT applications considering the user subscription plans and SLAs. In particular, mobile operators may dynamically program the network to address the traffic needs while considering UL and DL traffic separately, resolving situations that could lead to congestion. In addition, mobile operators can handle efficiently the service elasticity requirements of cloud providers enhancing the quality of experience taking full advantage of the network resource availability.

4.9 Conclusion

This section introduced an SDN-based network management architecture and control mechanisms to provide resource elasticity in a TD-LTE system. Such a resource management flexibility introduced by this proposal can enhance the UL/DL resource diversity in a RAN deployment increasing performance gains, while providing a key enabler for the network operators to support a broad range of OTT application and cloud services with a wide variety of UL/DL traffic demands. The proposed mechanism can address flexibly UL and DL traffic requirements in an autonomous manner addressing effectively pseudo-congestion by forming virtual cells enabling users to utilize resources from multiple eNBs allowing customized frames that resolve pseudo-congestion. An algorithm to manage the network resource providing flexibility and enhanced user performance has been introduced at the SDN controller to assess congestion situations and provide resolution via virtual cell provision and by enforcing TD-LTE frame configuration at selected eNBs. The results obtained via system level simulations show significant improvements in the average users’ performance including both edge users that reside within the virtual cell region and inner cell ones, without compromising the overall system performance. Further research is envisioned toward extending the proposed SDN based mechanisms considering split bearers and towards dynamic adjustments in the mobile backhaul provisioning resources for lower layer transport mechanisms.
5. Core Network Virtualization and Mobility Management (KCL)

5.1 Introduction

In this chapter of the deliverable D2.2 the main contribution throughout the second work package of CROSSFIRE project is presented. The objectives, as they were stated at the first deliverable D2.1 were met and the future aims are also given. In particular, in the first section an alternative approached based on the distributed mobility management logic for next generation wireless networks is presented. The performance benefit of the proposed scheme can be up to 17% in comparison to so far proposed distributed mobility management scheme by IETF. Moreover, in the second section a virtual network embedding algorithm that takes into account the users mobility effect is proposed and evaluated. The presented virtual network embedding algorithm outperforms existing algorithms from the literature in terms of routing cost as well as acceptance rate. Lastly, the future work objectives are given.

5.2 Hybrid DMM for Next-Generation Wireless Networks

One of key challenges for emerging and future wireless networks will be the support of seamless distributed IP mobility management to support a plethora of different applications. In this work, an optimization problem for provisioning efficient centralized Mobility Agents (MAs) deployment is formulated, as well as a realistic model is developed for the Distributed Mobility Management (DMM) scheme. These are subsequently compared in terms of various key characteristics, such as routing cost, delay and topology dependence. Then, an innovative Hybrid Distributed Mobility Management (HDMM) scheme is presented that provides improved network performance in terms of handover support for delay sensitive flows, compared to fully DMM schemes, which their performance can be strongly topology-dependent. The proposed scheme combines the centralized and distributed mobility support, depending on the network’s topology characteristics. A wide set of numerical investigations reveal the advantages of the DMM scheme over the centralized scheme for different network cases and detail the reasons why future networks tend to decentralize mobility management functionalities. Simulation results, also, show that the proposed HDMM scheme can significantly improve the network’s performance and the achieved QoS of the end-users, allowing seamless mobility support for delay intolerant over-the-top services.
5.2.1 Introduction

As the number of mobile Internet users tends to increase steadily and a dramatic growth of data traffic demand is expected to continue taking place into the next few years[77], all-IP based networks and IP based mobility support are the de-facto solutions, as they can offer to the users seamless mobility between heterogeneous wireless networks, without interrupting their service. In order to address the mobility support problem, many different mobility management schemes have been proposed so far, being mainly divided into two different categories: host-based and network-based mobility support solutions. In host-based, the Mobile Nodes (MNs) take actively part in the mobility management procedure, while in the network-based, certain net- work elements are deployed which are entirely responsible for the mobility support procedure.

The Internet Engineering Task Force (IETF) standards organization proposed Mobile IPv4 (MIPv4) [78] & Mobile IPv6 (MIPv6) [79] as host-based mobility support solutions for all-IP networks, where two IP addresses are assigned to the MN: the Home Address (HoA), which is the fixed address to identify the MN and the Care of Address (CoA), which indicates the current position (IP subnet) of the MN. The network entity called Home Agent (HA) handles the mapping of HoA and CoA and it is the node, where all the traffic towards and from the MN flows through.

Hierarchical Mobile IPv6 (HMIPv6) [80] is another mobility support solution proposed by IETF, as an extension to MIPv6, in order to improve network performance, in cases where frequent handovers take place and there is a long distance between the MNs and their home domains. HMIPv6 introduces the Mobility Anchor Point (MAP), which handles the mobility inside local domains, acting like a local HA, while MIPv6 handles mobility between different domains. Increased number of MAPs across the network tends to improve the routing of the flows, but, on the other hand, increases the signalling traffic.

In order to eliminate the signalling overhead and avoid the client’s IP mobility software installation, which is used in the host-based mobility management schemes, a network-based protocol for the mobility support has been proposed by IETF called Proxy Mobile IPv6 (PMIPv6)[81]. PMIPv6 uses network entities in order to handle the mobility, on behalf of the MN and, in this way it overcomes the need to add any software at the mobile users, excluding them from participation in any mobility-related signalling. When a MN is located within a PMIPv6 domain, the serving network gives it a unique home network prefix, although it has no awareness of this procedure, considering the PMIPv6 domain as the home network.
Then all the traffic towards the MN is tunnelled from the Local Mobility Anchor (LMA), an enhanced version of the HA in MIPv6, to the Mobility Access Gateway (MAG), which runs on the Access Router (AR), detecting the mobility of the MN and triggering the mobility-related signalling with the LMA.

However, the aforementioned mobility management solutions are strongly centralized, leading to some potential disadvantages that have been well documented in the literature, which might impact the overall performance of network. Firstly, there might be suboptimal routing of the flows, which means that in order for the path to include a mobility anchor, flows might follow unnecessarily longer paths. In addition, the anchor points might, also, become points of congestion, which can lead to higher latency, with detrimental effects to the Quality of Service (QoS). Moreover, when mobility anchoring is taking place within the network, there are issues with network robustness, since the failure of certain anchor points can lead to losing mobility support for a high number of MNs.

For these reasons, IETF’s working group proposed Distributed Mobility Management (DMM)[82], a decentralized solution with specific requirements[83], where the main concept is the distribution of the mobility anchors by locating them at the ARs of the network. In this way, each flow is being optimally routed, while the handovers are handled by tunnelling the flow to the next AR. Such a solution would be especially attractive for cases, where the call session time is less than the cell residence time, resulting in a very small number of handovers. But this might not be always the case, especially with the current trend of reducing cell radius in order to increase the capacity and address the problem of increased data demand and limited available bandwidth.

In this chapter, an optimization framework for the optimal number, location and selection of the Mobility Agents (MAs) across an access/core network is presented, where a centralized mobility management scheme is applied, by formulating it into a mathematical programming problem. Then, a model of the operation of the DMM scheme is developed and we evaluate its performance by comparing it with the derived optimized centralized mobility support scheme. Based on the above mentioned comparison, an Hybrid-DMM scheme is proposed, where delay-sensitive flows are handled by mobility anchors within the network, so that in case of handovers their performance is not deteriorate, compared to a pure DMM support, where the handover delay might be highly dependent on network topology. The main concept is that, when the delay for some specific flows exceeds a threshold, these flows anchor to mobility agents, which are placed on hierarchically higher nodes across the
network. To the best of our knowledge, this is the first time so far that a hybrid distributed-centralized mobility management scheme is proposed. Finally, we compare the performance of our proposed scheme with the so far presented schemes proposed by IETF, and the results show that there is a significant improvement gain on the achieved QoS for the end-users.

5.2.2 Selected Related Works

In this section there is a brief review of main related works to this study on mobility management for next generation networks. First, the authors in [84] present a framework on the optimal micro-mobility management in broadband access networks. Firstly, they analyse the impact of the MAs's location within the access network on the total routing and mobility overhead cost. Then, they formulate the problem of the optimal number and location of the MAs and, based on the simulations results, they achieve a significant improvement in the overall network performance. However, in their analysis, the flows are taken into account as total demands by the ARs and not individual flows and, also, the only possible handovers occurring are single-hop handovers.

In [85] Zuniga et al. give an extensive description of the DMM's framework, as it has been modified by the IETF and the 3GPP so far. First, the exact motivation of DMM is explained by listing the current deployed mobility management schemes' main drawbacks. Then, there is a review of the solutions for distribution of the mobility management that IETF and 3GPP standards organisations have proposed. The authors conclude by presenting some possible solutions for the evolution of the 3GPP’s Evolved Packet Core (EPC) and they end up pointing out that DMM is a suitable solution for the future mobile networks.

The authors in [86] propose a network-based DMM scheme, where the logical function of the LMA splits and the mobility routing is located jointly at each AR with a mobility client function. The main objective of their work is to optimize the routing of the handover flows as well as the new flows, in order to improve the overall network performance. The results from their simulations confirm that their proposed scheme achieves better packet-delivery cost, tunnelling cost and total cost than previously proposed schemes (D-PMIP), but has increased signalling cost.

In [87] T.X. Do and Y. Kim focus on vehicular scenarios and flat architectures and they propose D-NEMO, a distributed mobility management scheme, where all of the ARs have the mobility management functions of mobility anchors and PMIPv6's MAG and where a proxy
router is responsible for handling the registration of the MNs. The numerical analysis shows that the proposed scheme achieves better performance in terms of handover latency, in comparison to existing mobility management schemes.

Li Yi et al. in [88] present a study on the Internet flows duration, extracting some very useful results, based on real data and then they show how the duration of flows can affect the performance of the DMM scheme. Afterwards, by highlighting that only 3% of the Internet flows have a very long duration, they perform a comparison between the performance of the centralized and the distributed mobility management scheme. By this comparison they concluded that in general the DMM can be more efficient way to handle localized mobility. They, also, recommended that if the HoA is classified according to the application type, then the few flows with very long session durations could be handled by a centralized mobility management scheme.

![Figure 30](image)

**Figure 30** (a) Centralized scheme, (b) DMM scheme and (c) HDMM scheme.

5.2.3 **System Model**

5.2.3.1 **Centralized mobility management scheme**

For the first part of the numerical investigations, as presented hereafter, a model of an access/core wireless network is created, and the focus is on utilizing mobility management, where anchoring is taking place within the network, i.e. centralized schemes. The aim is to determine the optimal number, location and selection by the ARs of the MAs in the network. The types of networks, which are used, closely resemble wireless networks. For this reason, random planar tree-like graphs, varying from highly interconnected to sparse structures, are implemented, in order to test the topology dependency. The main goal is to simulate, as realistically as possible, the existent core networks of cellular networks. The network in Figure 30 (a) is an example of a binary tree (the simulations’ worst-case topology),
while the network in Figure 30 (b) is a tree-like network with dense interconnections between the intermediate the nodes. As the number of the interconnections increases, the routing options for the flows increase as well, offering a better potential routing cost.

We adopt a mathematical programming based solution approach for optimizing the process of mobility anchoring in wireless networks. To this end, the network can be modelled as a directed graph $G = (V, E)$, where $V$ represents the set of nodes and $E$ is the set of links. Let $F: E \rightarrow N$ be a function defining the cost of each link and $c: E \rightarrow N$ be a function, which defines the capacity of each link. We consider $K$ different accepted flows in the network. Without loss of generality, the data traffic originates from the source node $g = 1$ and flows through the MAs which serve the destination ARs. Each flow $k \in K$ can be expressed as a demand $d_k$ which originates from the source node towards one particular AR of the set $T$ of ARs (we suppose the more realistic networking case scenario, where flows are unsplittable).

Let $q$ denote the number of flows per AR. Every link between two routers has a constant weight so different paths across the network have a different impact in terms of delay or routing cost. Since all of the traffic is routed through the MAs, it is obvious that the number and the location of them are affecting dramatically the total cost. More specifically, the MAs can be nodes of congestion leading to performance deterioration in terms of QoS in the network and also they severely affect the performance of handovers in terms of latency. Let $C_m$ represent the capacity of the node, where MA $m$ is located.

In this scenario, we pre-calculate the shortest paths from the source node $g = 1$ through every possible MA $m$ towards an AR $j$, which is the destination node for the flow $k$. Let $P$ express the set of those paths. Considering that each commodity can use only one path, let $p_{km} \in P$ be the path for the $k^{th}$ flow, calculated from the source node to its destination AR $j$ through the MA $m$. In order to express the total cost which is produced by the data traffic, we consider that the path $p_{km}$ from the source node $g$ to an AR $j$ consists of two parts: the path from the source node to the MA $m$ and the path from the MA to the destination AR $j$, where an overhead is added (expressed as the percentage $o$), in order to represent the encapsulation. Taking into consideration that each link has a specific cost, the routing cost for each flow is proportional to the total number of hops. It is obvious that as the number of deployed MAs ($L$) is higher, there are more available paths and the total routing cost can be potentially lower. In addition, their location plays a significant role in each different case, depending on the given demand distribution at the ARs.
In order to capture the mobility of the users, we consider a handover matrix $H$, where $h_{kj} \in (0, 1)$ is the probability that the flow $k$ moves to the AR $j$ (with $h_{kk} = 0$. In our scenario, and without loss of generality, one and two-hop handovers can occur, so we consider the handover matrix $w$ where $w_{kij} \in (0,1)$ is the probability that the flow $k$ moves to AR $i$ through the intermediate AR $j$. Note that in the following formulation, all of the flows whose destination AR $j$ are anchored to the same MA. Also, when referring to the set of ARs $T$, without loss of generality, we refer to the set of destination nodes which is created by taking a single flow for each ARs (e.g. $T = \{k_i, k_j, \ldots, k_n\}$ where $k_i$'s destination is AR$_1$, $k_j$'s is AR$_2$ etc.).

$$M_m = \begin{cases} 
1, & \text{if an MA is located on node } m \\
0, & \text{otherwise} 
\end{cases}$$

$$x_{km} = \begin{cases} 
1, & \text{if flow } k \text{ is anchored to MA } m \\
0, & \text{otherwise} 
\end{cases}$$

$$G_{kjm} = \begin{cases} 
1, & \text{if flow } k \text{ and AR } j \text{ are supported by MA } m \\
0, & \text{otherwise} 
\end{cases}$$

The total routing cost $\Xi$ can be written as follows:

$$\Xi = \sum_{k \in K} \sum_{m \in J} d_k p_{km} x_{km}$$

When a flow $k$ moves to an AR $j$, then the flow is routed through the MA that supports the flows of the new AR. If this MA is different than the previous one (inter-MA handover), then we consider an extra cost $Z_{kj}$, which is added to the total mobility cost:
When a flow $k$ moves to an AR $j$, then the flow is routed through the MA that supports the flows of the new AR. If this MA is different than the previous one (inter-MA handover), then we consider an extra cost $Z_{kj}$, which is added to the total mobility cost:

$$Z_{kj} = \alpha X_{kj} + \omega$$

where $X_{kj}$ is the distance between the two ARs, $\alpha$ is a weight variable and $\omega$ is a constant, which represents the signalling cost to the home agent through the Internet. The total mobility cost $\Psi$, due to the total handovers, can be written as follows:

$$\Psi = \sum_{k \in K} \sum_{m \in J} \sum_{j \in T} (h_{kj}(d_{km} x_{jm} + Z_{kj}(x_{jm} - G_{kjm})) + \sum_{l \in T} w_{kjl}(d_{lm} x_{lm} P_{lm} + Z_{jl}(x_{lm} - G_{jlm}))$$

Finally, the total cost is the summation of the routing cost and the mobility cost: $\Xi + \Psi$. Below, we express the problem in a mathematical programming setting:

minimize $(\Xi + \Psi)$

subject to

$$\sum_m M_m \leq L \ (a)$$

$$x_{km} \leq M_m, \ \forall k \in K, m \in J \ (b)$$

$$\sum_k x_{km} d_k \leq M_m C_m, \ \forall m \in J \ (c)$$

$$\sum_k d_k - \sum_k \sum_m d_{km} x_{km} \leq 0, \ (d)$$

$$A(\sum_k x_{km} - q) \leq M_m - 1, \ \forall k \in K, m \in J \ (e)$$

$$G_{kjm} \leq x_{km}, \ \forall k \in K, m, j \in J \ (f)$$

$$x_{km} + x_{jm} - G_{kjm} \leq 1, \ \forall k \in K, m, j \in J \ (g)$$
\[ x_{km} + x_{jm} - G_{km} \leq 1, \forall k \in K, m, j \in J \quad (h) \]

\[ \sum_k x_{km} = 1, \forall k \in K \quad (i) \]

\[ \sum_j x_{jm} = q, j \quad (j) \]

where constraint (a) sets the maximum number of MAs that will be deployed, (b) ensures that if flow k will be anchored to an MA m, this MA has to be deployed, (c) is a capacity constraint in order not to exceed the maximum capacity of MAs m, (d) ensures that all of the demands have to be satisfied, (e) ensures that all of the flows with the same destination AR have to be anchored to the same MA m, (f) and (g) make sure that if and only if both ARs’ domains are served by the same MA, decision variable G has to be 1. Equality constraint (h) makes sure that each flow will be anchored to only one MA and (i) sets the number of flows per AR. Note that A \in N can be a relatively big natural number (e.g. A = 2q).

As stated in the previous section, the proposed mathematical programming setting is inspired by [8], which has been augmented to consider multiple flows and sessions, that can experience multiple handovers (one-hop and two-hop).

### 5.2.3.2 Distributed Mobility Management

For the DMM scheme scenario, the mobility management function is distributed at the ARs of the network. As a result, at first place, all of the flows are routed through the shortest paths. The routing cost \( R_k \) for each flow k is:

\[ R_k = d_k p_{gj} \]

where \( d_k \) is the data demand of the flow and \( p_{gj} \) is the shortest path from the source node g to the destination AR j. Regarding the handover procedure, when a flow migrates from an AR j to another AR i, the traffic is tunnelled to the new AR i Figure 30 (b), with an added overhead o. In this scenario, the possible handovers that can occur are, also, one-hop and two-hop handovers. The mobility of the users is expressed by the handover matrix H, where
$h_{kj} \in (0, 1), h_{kk} = 0$ represents the probability that a flow $k$ moves to an AR $j$, whether the new AR is one or two hops away. The cost caused by the mobility of the users for one flow $k$ is:

$$H_k = o h_{kj} d_k p_{ji}$$

where $p_{ji}$ is the shortest path from the AR $j$ to the AR $i$ and $o$ the overhead, expressed as the ratio of the packet length of the tunnelled flow to the initial one. Finally, the total cost for all of the flows is calculated as follow:

$$C = \sum_{k \in K} (R_k + H_k)$$

### 5.2.3.3 Hybrid Distributed Mobility Management

According to the results, which are presented analytically in the next section, the aforementioned mobility management schemes and especially the DMM scheme (for instance in heterogeneous networks, where there can be no direct connectivity between different access points) can suffer from strong topology dependence. This means that, although DMM is a promising mobility management solution for next-generation wireless networks, there can exist areas across the network where the migrating flows suffer from sub-optimal routing and strong latency. This can affect the QoS to the end users, especially when using delay-sensitive applications (i.e. VoIP).

Motivated by this, we propose an hybrid scheme, where the mobility function is distributed at the ARs, except for certain areas, where handover procedures cause high latency due to the underlying network topology characteristics. In these areas of the network, the mobility of the flows is supported by mobility agents which are located hierarchically higher across the network, as seen in Figure 30 (c). The number and location of those MAs can be found by solving the optimization framework we presented in the section above for those flows.

As shown in algorithm 1, we firstly calculate the total cost of every flow, considering that we have DMM scheme and we define a threshold $\gamma$, above which the distributed mobility support switches to a centralized mobility management. Afterwards, we group the flows that will switch from DMM support to centralized. Then, we solve the optimization framework, as it
was described above for these flows only, in order to find the optimal location and selection of MAs which will serve these flows. (in Figure 30c this occurs at the two ARs in the middle of the topology). Then, the mobility support for these flows is handled by MAs (in Figure 30c, the MA has been located at the source node). The cost for the flows that are supported by DMM scheme can be computed as in the section above.

### Algorithm 1: HYBRID–DMM SCHEME

1. **Step 1** Calculate the cost of every flow, using DMM
2. **Step 2** Find the flows for which the total cost is greater than a predefined threshold γ
3. **Step 3** Solve the integer programming problem, as described in 2.3.1, for those flows only
4. **Step 4** Anchor those flows to the mobility agents found by the optimization framework, using CMM. The rest will be supported by DMM scheme

### 5.2.4 Numerical Evaluation

An important aspect, which has to be considered when implementing a mobility management scheme, is how the physical topology of the network can affect the overall network performance. For this reason, we compare the centralized mobility management scheme to the DMM scheme, as it is described by IETF in [6]. In terms of topology dependence, two different types of network topologies are simulated; a sparse tree-like network (which incorporates the binary tree as the worst case in terms of available connectivity) and a highly interconnected tree-like network (Figure 30a,b accordingly). The optimization framework was solved using MATLAB’s optimization toolbox.

The networks used in these simulations are consisted by 256 nodes, the weights of the links randomly assigned using a uniform distribution with integer values that range from 8 to 12 and the capacities of the nodes are set to at least 20% of the total demand. Regarding the centralized scheme, two MAs are deployed, similar to the network shown Figure 30a and ω = 30, a = 1.1. The sparse network is a topology similar to Fig 2.1a and the dense network similar to Fig 2.1b. There are q = 10 flows per AR and the demands are randomly assigned with a uniform distribution ranging from 10 to 15 units.

As Figure 31 shows, the DMM scheme can be affected in a greater degree by topological characteristics of connectivity in the network (25% increase in the cost for the sparse topology), than the centralized mobility management scheme (11% increase of cost.
for sparse network). The reason is that, although the DMM scheme offers optimal routing, while flows are tunnelled to a neighbour AR, during the handover procedure, depending on the topology, they can be routed through unnecessarily longer paths, which relates to the connectivity of the network and the distance between the ARs.

In our simulations, we consider a traffic model where users can have sessions amortized in terms of number of handovers that can take place during the session time. In that respect, the baseline assumption is that sessions last for one or two handovers. In Figure 32 the total cost for centralized and DMM scheme in different mobility cases is presented. Each case is described from a pair of probabilities; the first one is the probability of one-hop handovers and the second one is the probability of two-hop handovers (e.g. the pair (0.7,0.4) in the x-axis means that the average probability for one-hop handovers is 70% and the average probability for two-hop handovers is 40%, for flows that have already had an one-hop handover). The simulation results confirm that the DMM outperforms the centralized mobility management scheme in terms of total cost in all evaluated cases, having up to 35% less total cost for a high mobility scenario.

Furthermore, although DMM performs better than the centralized scheme, as already shown, we focus not only on the overall performance, but, also, on the QoS for each particular flow. Figure 33 represents the total cost per group of flows with the same destination AR (in this figure, the network has 511 nodes and there are depicted 256 different ARs, numbering from left to right, as in Figure 30a). The results show that there are areas across the network where particular flows experience a total cost of up to 30% more than the network’s average. In addition, as it is shown, there are destinations nodes (ARs) where the flows destined to them appear to have a total cost greater than the depicted threshold γ (87%).

The motivation behind the Hybrid Distributed Mobility Management (HDMM) scheme is the seamless mobility for delay sensitive applications. Our main goal is to improve the QoS of those flows that have increased total cost, greater than a specific threshold γ, due to the network topology characteristics, as shown in Figure 33. In this simulation we set the threshold γ = 87% of maximum total cost. To the best of our knowledge this is the first hybrid distributed-centralized mobility management scheme, so we are interested in comparing the so far proposed DMM scheme, as it was presented in section above with our HDMM. To this end, the proposed HDMM scheme is implemented by utilizing the mobility support for those flows with cost greater than the average to the 2 MAs located at hierarchically higher in the
network (the nodes which have been found with the optimization framework). The performance of the DMM and the HDMM scheme is presented in Figure 34. As it is shown, the gain of the proposed HDMM scheme for the aforementioned flows can be up to 17%, depending on the probabilities of the occurring handovers (i.e. last case in Figure 34).

**Figure 31 Total cost for different topologies**

**Figure 32 Total cost for different mobility scenarios**
**Figure 33** Total cost for each AR domain

**Figure 34** Performance of DMM and HDMM for different mobility scenarios
5.2.5 Conclusions

Next-generation wireless networks will inevitably tend towards a decentralization of the mobility management, in order to achieve better handover performance, more efficient routing and generally improved network performance. Simulation results show that DMM achieves better overall performance than the centralized scheme in terms of total cost. In addition, both of these solutions present notable topology dependence, with the DMM scheme being more affected by the network topology in some cases. Based on these findings, a Hybrid-DMM scheme is proposed, where the mobility function is also distributed at the ARs, except specific areas, where data forwarding can lead to very high latency and routing cost; for this reason, in HDMM, these areas are supported by MAs that are located hierarchically higher across the network. Simulations results show that the proposed scheme can achieve a significant performance improvement especially for supporting time-critical over-the-top services.

5.3 Mobility Aware Virtual Network Embedding

Over the last years, network virtualization has become one of the most promising solutions for sustainability towards the on-going increase of data demand in mobile networks. The problem of efficiently forming a virtual network has been studied extensively during the past years and many different solutions have been proposed but these studies have mainly focused on wired networks. The main purpose of this study is to provide an optimization framework for optimal virtual network embedding by explicitly considering the effect of the actual user mobility, assuming a Distributed Mobility Management (DMM) scheme. In addition, service differentiation is introduced, giving higher priority to time-critical over-the-top (OTT) services compared to elastic Internet applications. The performance of the proposed framework is compared to greedy heuristics algorithms that are mobility agnostic and numerical investigations reveal that the effect of mobility has an important role to play in the design of virtual networks. Additionally, the mobility aware scheme can provide tangible gains in the overall performance compared with the previous proposed schemes that do not take into account the effect of user mobility.
5.3.1 Introduction

Mobile network operators worldwide have witnessed a dramatic growth of data demand, which requires further investment. At the same time, revenues are reaching a plateau due to a flat charging model. This tends to turn them into ‘dumb pipes’ for the application providers[89]. For this reason, following the same concept as in wired networks, the virtualization of resources in order to achieve efficient network sharing has been considered as a promising solution for next-generation networks, including 5G. It is worth pointing out that some form of passive (non-adaptive) infrastructure sharing exists already within the cellular operators and it is expected that by the end of 2015 90% of mobile operators will have explored this avenue in some form\(^5\).

Clearly, this trend is expected to further continue in the future, but sharing will move deeper into the network. This will give the flexibility of more dynamic sharing via virtualization allowing for multi-tenancy at different network elements within the core and wireless access network. The main advantages of network virtualization are the increase of utilization of the available physical resources, the energy efficiency, the flexibility and scalability that offers and finally the prospect of sustainability for mobile operators [90].

In network virtualization, the physical resources (e.g. nodes and links) have to be virtualized in order to form virtual networks, respectively to virtual network requests, in a procedure named virtual network embedding. The problem of virtual network embedding has been studied extensively and a plethora of different approaches have been proposed. A very useful review of the so far research in the virtual network embedding area is presented in [91]. The authors index the most important optimization algorithms. Then, they categorize the algorithms along three main dimensions: static-dynamic, centralized-distributed and concise-redundant.

An important issue that has to be taken into account while developing a virtual network embedding algorithm for mobile networks is the effect of user mobility. The way that the mobility support is implemented (i.e. the mobility management scheme that is in use) affects the core network’s congestion, performance and available resources and subsequently the optimal virtual network set up.

Since the proposal of Mobile IPv6 (MIPv6) [79]by the Internet Engineering Task Force (IETF), various all-IP mobility management schemes have been standardized so far. The

modern trend is to move from centralization to the distribution of the mobility function at the edge of the core network. To this end, after setting the requirements [83], IETF proposed the Distributed Mobility Management scheme [82], where each edge router becomes a mobility anchor for the users. In DMM, during a handover, the flows are being tunneled directly to the new edge router, which is connected to the new base-station. Hereafter, we will assume that mobility management is handled in a distributed manner using a DMM compliant solution, where DMM mobility anchor points are located on the edge routers.

In this work we consider a mobile core network as the available physical infrastructure, which has to be virtualized in order different virtual networks to be set up, according to virtual network requests. First we consider different tenants that share the physical infrastructure and we provide tenant prioritization. An integer mathematical programming formulation is developed, which also takes into account the mobility of the users. Then, a greedy heuristic algorithm, which is mobility agnostic is presented and its performance is compared to the solution of the optimization framework. The main contribution of this proposal is that it reveals the interaction of physical network topology and traffic tunnelling at the mobility anchors on the efficient construction of virtual networks. To this end, the proposed integer programming formulation considers in an explicit way the users’ mobility and the effect of that mobility on network resources. Moreover, it performs prioritization among multiple tenants with potential differentiation in terms of QoS (but the model can be easily extended to differentiate charging model and QoE).

In Figure 35 an example of the outcome of a mobility agnostic and a mobility aware virtual embedding algorithm respectively is presented. Firstly, we consider a substrate network where a virtual network has to be formed in order to serve a class of flows, a fraction of which moves between the two edge routers as depicted. The handover is handled by DMM scheme and the intermediate path between the two edge routers is used (e.g. in Evolved Packet Core (EPC) architecture these could be the serving gateways). Then, as shown in Figure 30a, given the gateway and the destination edge router, which is defined by the virtual network request, the mobility agnostic algorithm will not include the essential for the handover procedure tunnelling path in the virtual network. In this way, the DMM tunnelling will have to be routed through a path consisted of five intermediate hops.

On the other hand, the proposed mobility aware algorithm takes into consideration not only the source and destination nodes but also the DMM tunnelling path for the fraction of the flows that move. Hence, the algorithm will firstly map the two shortest paths that connect the
gateway with the base-stations (for flows that start from there) and then it will additionally include the intermediate path as shown in Figure 35b. In this way, the DMMtunnelling path, connecting the previous DMM edge router with the new one, will be improved, being consisted by three hops.

![Figure 35](image)

**Figure 35** (a) Mobility agnostic and (b) mobility aware virtual network embedding algorithms

The key difference between the two approaches is that an embedding algorithm, which explicitly utilizes mobility information, will map an extra path that will be used for the data tunnelling to the next mobility anchor point, where the flow will migrate. Considering the users mobility effect, mobility aware algorithm will create virtual networks that can support migrating flows more efficiently. To the best of our knowledge, this is the first time a mobility aware optimal virtual network embedding algorithm is developed.

### 5.3.2 Previous related work

Different approaches have been proposed so far that address the virtual network embedding problem. In this section we include selected related works from the literature. The authors in [92] formulate the virtual embedding problem as a mixed integer program and then they propose virtual embedding algorithms by introducing a coordination between node and
link mapping phases. More specifically, they relax the integer constraints to get a linear programming setting by using deterministic and randomized rounding techniques and in this way they manage to achieve polynomial-time albeit suboptimal solutions for virtual network embedding. The simulation results show that their algorithms outperform the existing approaches in terms of acceptance ratio, revenue and provisioning cost.

The authors in [93] propose an integer-linear programming formulation that solves the on-line virtual network embedding problem. Their solution aims to minimize the total resource consumption and to achieve load balancing across the network. To this end, they introduce three cost functions: one that minimizes the total load on each virtual network, one that minimizes the number of links that are mapped and selects nodes with higher available resources and one that includes the demanded capacity by the virtual network requests in the objective function. The simulation results show that their proposal outperforms in general the different compared heuristic approaches.

A virtual network embedding distributed protocol, named MADE suitable for mobile environments, where the nodes are not static and the substrate network is dynamic, is presented in [94]. The authors apply the path splitting and migration techniques in order to optimize the utilization of the available physical resources. The simulation results show the efficiency of this protocol in terms of acceptance and completion of the requested virtual networks.

In [95] a distributed algorithm that performs load balancing and virtual network embedding over a substrate network is presented. The algorithm makes use of a proposed mapping protocol that enables the communication and the exchange of messages among the substrate network’s nodes in order to ensure distributed negotiation and synchronization for the virtual network set up. The performance results show that in the distributed mapping approach the number of messages exchanged may have an important impact on the overall performance but, on the other hand, compared to the centralized approach, their proposed algorithm can reduce the time delay and it can manage to process multiple parallel virtual network requests.

The authors in [96] develop a scalable embedding algorithm named VNE-AC, based on ant colony metaheuristic. The algorithm aims to minimize the allocated physical resources of the physical substrate network for each request in order to minimize the reject rate and to maximize the provider’s revenue. Results from the simulations show that the proposed algorithm achieves better overall performance in comparison to related algorithms from the
literature. In [97] is presented an alternative approach for the virtual embedding problem that focuses on rethinking the designing of the substrate network per se, in order to enable less complex algorithms and increase the utilization of the resources, without the restriction of the problem space. In order to achieve flexibility of the substrate network, path splitting and migration as well as customized node-mapping algorithms are used. The simulations results show that the proposed solution is competitive and it manages to increase the resources' utilization.

The authors in [98] focus on solving the problem of virtual network embedding while the substrate network evolves. To this end, they present an integer programming formulation of this problem that aims to minimize the upgrading cost of virtual network with respect to node resource and path delay constraints. Because of the problem’s complexity the authors develop a heuristic algorithm and they present its efficiency through simulations.

In all the above previous research works the effect of users mobility has not been considered. As shown in the sequel, this is an important parameter that needs to be taken into account for creating efficient virtual networks.

5.3.3 System Model Description

5.3.3.1 Implementation of virtual network embedding algorithms

We consider that in order the virtual network requests to be served, a central controller exists, being responsible for slicing the physical resources and also providing isolation among them. In particular, the controller is handled by a virtual network embedding algorithm, which means that the embedding algorithm creates a plan/strategy for virtualization of the resources and for the assignment to the different requests and then the controller applies it. Each algorithm has different strategy and logic in order to achieve efficient mapping of the physical resources, depending on the parameters that takes into consideration.

After the decision for the assignment of the resources, the controller is responsible for communicating with the network elements and creating the slices. In this way, virtual networks are formed, ready to serve the flows that they are responsible to handle. Each slice is totally isolated from the other slices and has no awareness of their existence.

Regarding the architecture, for our topology, we consider a single gateway and several destination edge routers. Each virtual network request is defined by the demands
from the source gateway to the end edge routers. Hence, the formed virtual networks have a substrate topology dependence and should include the same source and destination nodes.

As for the mobility management function, since DMM scheme is deployed, every single destination edge router acts as a mobility management anchor point. Hence, we consider that by applying Software Defined Networking (SDN) logic and Network Function Virtualization (NFV) deployment, mobility functions are virtualized and activated on demand at the edge of the core network (i.e. the edge routers).

5.3.3.2 Mathematical programming setting

In this section we present the proposed mathematical programming setting for optimal mobility aware virtual network embedding. In order to provide a formal model for the virtual network embedding process and in order to develop the optimization framework, firstly we detail the physical substrate network, then the requests for virtual networks, the objective function and the problem constraints.

**Substrate network:** the substrate network can be modelled as an undirected tree-like graph $G = (V, E)$, where set $V$ represents the set of nodes and set $E$ depicts the set of links. Let $B: E \rightarrow \mathbb{R} > 0$ be the cost of each link and $C: V \rightarrow \mathbb{R} > 0$ the capacity of each node. The gateway is located on the source node of the graph, while the leaves of the tree represent the destination edge routers. The intermediate nodes are the routers of the core network (Figure 36, Figure 37). The algorithm will assign nodes and links according to the virtual network requests.

- Virtual network requests: the proposed algorithm performs a multi-tenant prioritization. This can be applied in a case where the infrastructure owner provides a
tenant differentiation according to desirable QoS requirements. This, also, applies to a scenario where a prioritization in favor of one type of service over another one with lower priority is demanded. For this reason we will refer to this feature of our proposed algorithm as tenant or type of service classification/differentiation. Let \( Q \) represent the set of sets of the virtual network requests for all tenants. Moreover, let \( U \) represent the set of the different tenants. Then, let \( Q_u \in U \) represent the set of virtual network requests for a tenant \( u \in U \). The set \( Q \) can be described as follows: \( Q = U \cup Q_u \cup ... \cup Q_U \).

In order to define the set of virtual network requests \( Q_u \in Q \) we consider that each request \( q \in Q_u \) is described by a set of classes of unsplittable flows \( k \in K_u \) that have to be satisfied and routed from the root gateway to one edge router \( j \in T \) by the formed virtual networks. Also, each class of flows \( k \in K_u \) corresponds to a total demand \( d_{uq} \). Hence the virtual network request \( q \in Q_u \) is defined by a set of two nodes (source-destination nodes) undirected graphs - paths that can be notated as \( \pi_{kp} \in P \). Each class of flow of a virtual network request can be considered as a demand for resource allocation across the core network. We hereafter explicitly consider a per-class allocation of the available network resources allowing for scalability and assuming that requested traffic demand incorporates sufficient slack values, so that traffic variations (congestion episodes) during the duration of the virtual network are taken into account.

- Mobility: in order to capture the actual mobility of the users, we consider a handover matrix \( H_{K \times K} \), the elements \( h_{kj} \in (0,1) \) of which represent the probability of the flow \( k \in K_u \) to migrate to another edge router \( j \in T \) (note that each edge router \( j \) can be described by a flow \( k \in K_u \)). When DMM scheme is in use, there is an additional need to assign paths that connect edge routers or network elements where DMM anchoring is taking place.
In this work and without loss of generality we assume that DMM anchoring is taking place at the edge router and for this reason, for each edge router of class of flow \( k \in K_{uq} \) towards an edge router \( j \in T \) we consider the set of substrate paths \( r_{kj} \in R \). In the same way, each tunnelled class of flows will be routed among one of the available paths. However, our optimization algorithm can be adapted for different other mobility management solutions, since mobility cost depends on the mobility management scheme that is used. In conclusion, the algorithm, taking into account the probabilities for handovers, will also map paths for the virtual networks for more efficient data forwarding.

- Problem variables: based on the above setting, the goal is to find the optimal selection of routing paths in order to minimize the total routing cost and at the same time to achieve tenant differentiation. Below, there is a summary of the variables that used for the formulation of the integer mathematical program and the introduction of the decision variables:

\[ G = (V, E): \text{undirected planar tree-like graph} \]
\[ \pi_{kp}: \text{set of substrate paths from source to edge router } k \]
\[ r_{kj}: \text{set of substrate paths from edge router defined by flow } k \in K_{uq} \text{ to edge router } j \in T \]
\[ U: \text{set of different tenant’s set of virtual network requests} \]
\[ Q_u: \text{set of virtual network requests of tenant } u \in U \]
\[ K_{uq}: \text{set of classes of flows that have to be served by virtual network request } q \in Q_u \]
\[ d_{uq}: \text{set of demands for class of flows } k \in K_{uq} \text{ of tenant } u \in U \text{ and virtual } k \text{ network request } q \in Q_u \]

\[
z_{kp}^n = \begin{cases} 
1, & \text{if node } n \in \pi_{kp} \\
0, & \text{otherwise}
\end{cases}
\]
\[
r_{kj}^n = \begin{cases} 
1, & \text{if node } n \in r_{kj} \\
0, & \text{otherwise}
\end{cases}
\]

Then, we define the following Boolean variables:
\[ x_{kp}^{uq} = \begin{cases} 1, & \text{if } \pi_{kp} \text{ is assigned to } k \in K^{uq} \\ 0, & \text{otherwise} \end{cases} \]

\[ y_{kji}^{uq} = \begin{cases} 1, & \text{if } r_{kji} \text{ is assigned to } k \in K^{uq} \\ 0, & \text{otherwise} \end{cases} \]

- Objective function and problem constraints

The total routing cost \( \varphi \) can be written as:

\[ \varphi = \sum_{u,q,k,p} d_{k}^{uq} \pi_{kp} x_{kp}^{uq} \]

The total mobility cost \( M \) can be written as:

\[ M = \sum_{u,q,k,j} h_{k}^{d} r_{kji} y_{kji}^{uq} \]

If \( T = M + \varphi \) then based on the above definitions the mathematical program can be formulated as follows:

minimize \( T \)

subject to

\[ \sum_{u,q,k} \left( \sum_{p} d_{k}^{uq} x_{kp}^{uq} + \sum_{i,j} h_{k}^{d} r_{kji} y_{kji}^{uq} \right) \leq C \quad \forall \ n \in V \quad (a) \]

\[ T_{un} \leq T_{un+1} \quad \forall \ u, q, k, p, i \quad (b) \]

\[ \sum_{j} y_{kji}^{uq} \leq \vartheta(h_{k}) \quad \forall \ u, q, k \quad (c) \]

\[ \sum_{p} x_{kp}^{uq} = 1 \quad \forall \ u, q, k \quad (d) \]

\[ x_{kp}^{uq}, y_{kji}^{uq} \in \{0,1\} \quad \forall \ u, q, k, p, i \quad (e) \]
where constraint (a) makes sure that the capacity of each node is not violated. Constraint (b) ensures the classification of the two types of service. Moreover, constraint (c) guarantees that if there is a handover between two ARs (defined by matrix h, then one and only one path will be used. Note that l is an indicator function defined as follows:

\[
\vartheta(n) = \begin{cases} 
1, & \text{if } n \neq 0 \\
0, & \text{otherwise}
\end{cases}
\]

Finally, constraint (d) ensures that among all the alternative paths which connect the source node with an edge router k only one will be used and constraint (e) makes sure that the decision variables will be Boolean. Since integer linear programming is NP-complete [99], problem instances related to large networks might be intractable and so low complexity heuristic methods are useful for such instances. A mobility agnostic greedy heuristic algorithm is presented below, but we note that since these types of optimization problems do not require real time operation, optimal solutions can be achieved.

5.3.3.3 Greedy Heuristic Algorithm

For the heuristic approach of the above problem, we implement a generalized greedy algorithm similar to a heuristic algorithm presented in [97], by augmenting to multiple traffic classes. The algorithm firstly sorts the demands of every virtual network request of each type of service. Then, it serves first the high-priority tenant in the following way: the flows with the higher aggregate demands utilize the lower-cost routing paths, until a threshold on the congestion level of these paths is reached, and then the flows with the lower aggregate demands follow. After the high-priority service has been assigned with routing paths, the algorithm serves the lower priority tenant.
The steps of this algorithm are presented in Alg. 2. Note that the mobility agnostic greedy algorithm embeds networks where the mobility is not taken into account. In this way, the tunnelling of the flows from the previous to the new DMM edge routers will use the mapped paths that connect the gateway with the edge nodes (as we already explained at the exampled shown in Figure 35). The total cost is the summation of the routing plus the mobility cost for every class of flow $k$.

Regarding the construction of virtual networks, it is not required to consider it as a network function with on-line service requirements. Therefore, the proposed set of mechanisms are envisaged to take place in a pseudo-real time manner. Also, in terms of network provisioning, a virtual network-enabled cloud environment is considered; hence, some form of centralization as required for the proposed optimization problem can be available. It needs to be further noted that as more network functionalities with diverse set of characteristics and performance requirements move to such cloud-based environment, focus should be turned on an efficient processing of the control plane. These issues clearly relate to the network performance as a whole but fall beyond the scope of this work.
5.3.4 Numerical Investigations

The integer programming problem was solved using the optimization toolbox of MATLAB. As we have already mentioned previously, the key innovation of the proposed algorithm in comparison to existing virtual embedding algorithms from the literature is that we have modelled the effect of users’ mobility and this has been taken explicitly into consideration. In order to capture the effect of mobility on the virtual network embedding process we first solve the proposed algorithm and then we compare the results to the performance of the mobility agnostic heuristic algorithm.

The impact of mobility is presented in

![Graph showing the impact of mobility on total cost](image)

Figure 38. The percentage s denotes the fraction of the flows that migrate to neighbour edge routers due to the mobility of the users. For s = 0% there is no mobility and there is a fixed network case. On the other hand, when s = 100% every flow moves to another domain. The topology that was used is the one shown in Figure 37. In particular, in this plot it can be concluded that as the mobility increases, the total cost of the proposed optimization framework outperforms the greedy algorithm that doesn’t take into consideration the mobility of the users. When there is no mobility the two algorithms perform the same. This is due to the fact that the parameters’ set that was applied led to a low congestion, making the solution by the greedy algorithm to be the same as the optimal. This means that in a low congestion case with no mobility, the heuristic performs very well. However, when
50% of the flows migrate, the proposed algorithm achieves 38% less total cost. In the upper bound scenario, for $s = 100\%$, the proposed algorithm achieves a 47% total cost reduction compared to the mobility agnostic algorithm.

Then, we evaluate the performance of the integer programming algorithm, as described in the above section, in terms of the acceptance rate. The acceptance rate is a metric that expresses the probability of successfully serving every virtual network request for a specific scenario. In cases where the capacity of the substrate network is much greater than the total demand, the mapping of each virtual network is plainly the set of the optimal paths, since the least cost paths can afford the requested aggregate traffic. However, this is not always the case. In scenarios where the capacity of the network can be comparable to the total demand of the virtual network requests, there might occur a high utilization of the resources and a congestion episode has a high probability to take place. In these cases, the optimization framework cannot always serve all of the virtual network requests. For this reason, in order to evaluate the acceptance rate, we compute the outcome of the above presented algorithms by running them multiple times, inputting the random variables.

A crucial parameter that affects the acceptance rate of the optimization algorithm is the number of the alternative paths $P$ and $R$ for every pair of source-destination in the network. This number is used to calculate the set of available paths among which the algorithm forms the virtual networks. As this number increases, the algorithm has a larger set of possible solutions and in this way it makes the successful forming of virtual networks embedding more possible to take place.

As shown in Figure 39, regarding the acceptance rate of the integer program for the above scenario, as the number of the alternative paths $P, R$ increases, the algorithm has indeed a higher probability of forming successfully the incoming virtual network requests. In particular, when the number of alternative shortest paths is 2, the acceptance rate is 31%, while when the alternative paths are 10, the acceptance rate increases by 42%.

**In order to compare the optimality gap of the solution of the proposed integer programming algorithm and of the greedy heuristic algorithm, we add a constant**
random variable of the nodes’ capacity. For the scenario with the aforementioned parameter values, as Figure 40 shows, the optimization algorithm achieves much higher acceptance rate than the greedy heuristic. In particular, in cases where the greedy algorithm cannot serve the virtual network requests, the optimization algorithm has an acceptance rate up to 50%. This means that in crucial cases, where the utilization of the physical resources increases, the optimization algorithm manages to deal better with the virtual network requests than the greedy heuristic algorithm.

Figure 36 Dense tree topology modelling a substrate network with 31 nodes
The reason behind that is the fact that the greedy algorithm sorts the demands of each class of flows for every type of service and then, serially assigns the best paths to the biggest demands. In this way, due to the fact that while mapping, it doesn’t consider the whole set of the demands, it possibly congests some of the paths and subsequently it fails to set up the virtual network according to the requests.
Figure 39 Acceptance rate vs. number of different alternative paths

Figure 40 Acceptance rate vs. mean node capacity value
This resembles to the optimality gap of a bin-packing problem. In addition, the algorithm tends to congest more the network than the optimization algorithm (Figure 38). On the other hand, the integer program finds the optimal solution for the whole network, taking into account every virtual network request at the same time, when a feasible solution exists. Lastly, we capture the impact of the network topology.
on the total cost, as it is shown above. In particular, we firstly solve the integer programming algorithm using a dense-tree like topology (Figure 36) as with the above cases. Then, we solve the algorithm using the same parameter values but we change the topology to a sparse-tree like network (Figure 37). The impact of the network topology is notable, reaching for this scenario an increase of the total cost of about 16% (Figure 41).

5.3.5 Conclusions

With the ever-upcoming pressure to increase overall network capacity, while offering almost flat rates to customers for mobile Internet access network, sharing has become a vital
component to achieve network sustainability. The expectation is that network sharing via virtualization will move deeper into the core network of operators and will allow for significant lower capital and operational expenditure. A key technical aspect of network virtualization is virtual network embedding, which relates to efficient mapping between physical and virtual resources. The problem of virtual network embedding has been previously studied, mainly for fixed networks, without taking into account the effect of mobility management solution, which affects routing decisions in case of user mobility. In this work, we reveal the impact of users' mobility on the virtual network embedding procedure and how we might entail in suboptimal virtual networks if mobility is not considered. To this end, we formulate the problem of virtual network embedding as an integer mathematical program in a way that takes explicitly into account the users mobility effect. The simulation results show that the mobility of the users has a significant impact on the virtual network embedding procedure. Numerical investigations reveal that the performance and efficiency of the network can be significantly increased when the proposed mobility aware algorithm is compared with embedding algorithms that are mobility agnostic.

5.4 Future Steps

The research about this work-package’s part is going to move deeper into virtual network embedding as well as network sharing techniques. Already, two more works on different approaches on the virtual network embedding have been made and submitted to major IEEE conferences.

Moreover, collaboration between the industrial partner Steinwurf ApS in Aalborg, Denmark is expected to take place during the secondment (1st of May, 31st of July 2015). During this period the objective is to produce a joint work on network virtualization and network coding.

6. Conclusions

This deliverable provides the initial output of CROSSFIRE in the field of network virtualization for mobile infrastructures concentrating of LTE-A. In particular it introduces a network architecture and an algorithm called RENEV for managing the resources of LTE-A considering a heterogeneous scenario, composed of an eNB overlaid with small cells. This scheme considers the coordination among several BSs to create an abstraction of systems’
radio resources, so that multiple tenants can be served. The extensive performance assessment has revealed that gains in system throughput are translated into gains for the users’ perspective. Finally the solution has been evaluated for the signaling overhead that adds into the network for increasing number of SCs per cluster. The proposed analytical formulation can provide useful insights that can be exploited for effective network planning of SCs in realistic heterogeneous environments towards 5G virtualized architectures. In addition, it details an SDN architecture and an Open Flow based protocol for providing on-demand radio and network resource sharing among different mobile network operators at selected base stations, called open eNBs, that act as multi-tenant base stations. The proposed architecture can enable the mobile network operators to bring the access network closer to their users, by utilizing the base station infrastructure of other operators. Moreover it launches the design of innovative sharing business models in the future cellular network, while it can also be used to offload the traffic from a tagged network operator (home operator) to the base stations of another network operator (host operator) under high duty cycles, e.g. crowded public events.

Furthermore, this deliverable introduces the concept of virtual cells in TD-LTE networks enabling a customized resource usage considering the resource demands of particular applications. Virtual cells with the help of SDN based network management and control mechanisms provide flexibility and allows network operators to adaptively manage the network as per the OTT applications’ or services’ requirements. This enables the network operators to efficiently support wide range of OTT applications and services with diverse UL/DL requirements. Finally, it introduces core network virtualization elaborating a network hybrid distributed mobility management and a virtual network embedding mobility-aware solution. The main contribution of the mobility management proposal is that facilitates seamless mobility by considering the network’s topology characteristics. On the other hand, the proposed virtual network embedding solution explicitly takes into account the users’ mobility effect allowing for more efficient network multi-tenancy.
### List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>Third Generation Partnership Project</td>
</tr>
<tr>
<td>API</td>
<td>Application Program Interface</td>
</tr>
<tr>
<td>AS</td>
<td>Access Stratum</td>
</tr>
<tr>
<td>AWGN</td>
<td>Additive White Gaussian Noise</td>
</tr>
<tr>
<td>BGP</td>
<td>Border Gateway Protocol</td>
</tr>
<tr>
<td>BS</td>
<td>Base Station</td>
</tr>
<tr>
<td>A-CPI</td>
<td>Application- Controller Plane Interfaces</td>
</tr>
<tr>
<td>D-CPI</td>
<td>Data-Controller Plane Interfaces</td>
</tr>
<tr>
<td>eNB</td>
<td>evolved NodeB</td>
</tr>
<tr>
<td>EPC</td>
<td>Evolved Packet Core</td>
</tr>
<tr>
<td>E-UTRA</td>
<td>Evolved Universal Terrestrial Radio Access</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>HeNB GW</td>
<td>Home eNB Gateway</td>
</tr>
<tr>
<td>HeNBs</td>
<td>Home eNBs</td>
</tr>
<tr>
<td>HetNets</td>
<td>Heterogeneous Networks</td>
</tr>
<tr>
<td>ICIC</td>
<td>Inter-Cell Interference Coordination</td>
</tr>
<tr>
<td>LTE-A</td>
<td>Long Term Evolution-Advanced</td>
</tr>
<tr>
<td>MAC</td>
<td>Medium Access Control</td>
</tr>
<tr>
<td>MCS</td>
<td>Modulation and Coding Scheme</td>
</tr>
<tr>
<td>MME</td>
<td>Mobility Management Entity</td>
</tr>
<tr>
<td>MPLS</td>
<td>Multi-protocol Label Switching</td>
</tr>
<tr>
<td>NVS</td>
<td>Network Virtualization Substrate</td>
</tr>
<tr>
<td>OFDM</td>
<td>Orthogonal Frequency Division</td>
</tr>
<tr>
<td>OFDMA</td>
<td>Orthogonal Frequency Division Multiple Access</td>
</tr>
<tr>
<td>PRR</td>
<td>Partial Resource Reservation</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RAC</td>
<td>Radio Admission Control</td>
</tr>
<tr>
<td>RAN</td>
<td>Radio Access Network</td>
</tr>
<tr>
<td>RBC</td>
<td>Radio Bearer Control</td>
</tr>
<tr>
<td>Rs</td>
<td>Resource Blocks</td>
</tr>
<tr>
<td>RENEV</td>
<td>Resources nEgotiation for NEtwork Virtualization</td>
</tr>
<tr>
<td>RRC</td>
<td>Radio Resource Control</td>
</tr>
<tr>
<td>RRM</td>
<td>Radio Resource Management</td>
</tr>
<tr>
<td>SCs</td>
<td>Small Cells</td>
</tr>
<tr>
<td>SISO</td>
<td>Single Input Single Output</td>
</tr>
<tr>
<td>SLA</td>
<td>Service Level Agreement</td>
</tr>
<tr>
<td>SDN</td>
<td>Software Defined Network</td>
</tr>
<tr>
<td>TDD</td>
<td>Time Division Duplex</td>
</tr>
<tr>
<td>UE</td>
<td>User Equipment</td>
</tr>
<tr>
<td>VPN</td>
<td>Virtual Private Networks</td>
</tr>
</tbody>
</table>
References

[1]. 3GPP, Overview of 3GPP Release 11 V0.1.4, 3GPP Std., March 2013.
[4]. 3GPP TS 36.331, Radio Resource Control (RRC); Protocol specification, Std., Rev. 11.4.0, 2013.
[12]. 3GPP, Overview of 3GPP Release 11, 3GPP Std., Rev. v0.1.7, December 2013.


[29]. 3GPP TR 36.913, Requirements for further advancements for E-UTRA, 3GPP Std., Rev. 11.0.0, September 2012.


[31]. 3GPP TS 36.412, E-UTRAN S1 signalling transport, 3GPP Std., Rev. 12.0.0, September 2014.


[33]. 3GPP TS 36.331, Radio Resource Control (RRC); Protocol specification, 3GPP Std., Rev. 12.3.0, September 2014.
[35]. 3GPP TS 36.413, E-UTRAN S1 Application Protocol (S1AP), 3GPP Std., Septemember 2013.


