Traffic and Mobility Management in Large-Scale Networks of Small Cells

Ph.D. Thesis Dissertation

by

Jaime Ferragut Martínez-Vara de Rey

Submitted to the Universitat Politècnica de Catalunya – BarcelonaTech

in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Barcelona, June 2014

Supervised by Dr. Josep Mangues-Bafalluy

Ph.D. Program in Telematics Engineering

Acknowledgements

To Josep Mangues and Miquel Soriano, for their constant support and invaluable advice. To Gil Zussman and all his team at Columbia University, for giving me the opportunity to enjoy one of the most fulfilling professional and personal experiences in my life. To all my colleagues at CTTC, for the good moments we have shared during these five years. To my friends and family, for their infinite patience and unconditional support. To Sarah, for being by my side during all these years (and many more to come!). And, especially, to my dad, for teaching me the most important lessons in life.

This work has been partially supported by Spanish grants TEC2011-29700-C02-01, FPU AP2009-5000, and European Union grant ICT-4-248523.

Contents

List of Figures

List of Tables

List of Acronyms

3GPP	Third-Generation Partnership Project
ACK	Acknowledgement
ADSL	Asymmetric Digital Subscriber Line
AIR	Antenna-Integrated Radio
AMuSe	Adaptive Multicast Service
AP	Access Point
API	Application Programming Interface
APN	Access Point Name
ARP	Address Resolution Protocol
ARP	Allocation and Retention Priority
ARQ	Automatic Repeat Request
ASN.1	Abstract Syntax Notation One
AuC	Authentication Centre
AWGN	Additive White Gaussian Noise
BBU	Baseband Unit
BeFEMTO	Broadband Evolved Femto Networks
BM-SC	Broadcast/Multicast Service Centre
BPSK	Binary Phase-Shift Keying
C-RAN	Cloud RAN
C-RAN	Centralised RAN
CA	Carrier Allocation
CAPEX	Capital Expenditure
CDMA2000	Code Division Multiple Access 2000
CPRI	Common Public Radio Interface
CS	Circuit Switched
CSG	Closed Subscriber Group
D-bearer	Direct bearer
D2D	Device-to-Device
DAS	Distributed Antenna System
DASH	Dynamic Adaptive Streaming over HTTP
DiffServ	Differentiated Services

DL	Downlink
DSCP	Differentiated Services Code Point
DSP	Digital Signal Processor
DTMC	Discrete-Time Markov Chain
E-RAB	Evolved Radio Access Bearer
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
eMBMS	Evolved Multicast/Broadcast Multimedia Service
eNodeB	Evolved Node B
EPC	Evolved Packet Core
EPS	Evolved Packet System
FB	Feedback
FBN-LIST	Feedback Nodes List
FDP	Fault Diagnosis Probe
FEC	Forward Error Correction
FIFO	First-In First-Out
FM	Fault Management
FPR	False Positive Ratio
GBR	Guaranteed Bit Rate
GPRS	General Packet Radio Service
GPS	Global Positioning System
GTP-C	GPRS Tunnelling Protocol Control Plane
GTP-U	GPRS Tunnelling Protocol User Plane
H-Bearer	Inter-HeNB Bearer
HARQ	Hybrid Automatic Repeat Request
HeMS	Home Evolved Node B Management System
HeNB	Home Evolved Node B
HeNB GW	Home eNB Gateway
HetNet	Heterogeneous Network
HO	Handover
HSS	Home Subscriber Server
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
ICT	Information and Communication Technologies

IE	Information Element
IETF	Internet Engineering Task Force
IFOM	IP Flow Mobility
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPSec	Internet Protocol Security
ISP	Internet Service Provider
ITU-T	International Telecommunications Union, Standardisation Sector
L-GW	Local Gateway
LBP	Leader-Based Protocol
LIMONET	LIPA Mobility and SIPTO at the Local Network
LIPA	Local IP Access
LLM	Local Location Management
LNGW	Local Network of Small Cells Gateway
LNM	Local Network Manager
LNO	Local Network Operator
LQ	Link Quality
LTE	Long-Term Evolution
M2-AP	M2 Application Protocol
M3-AP	M3 Application Protocol
MAC	Medium Access Control
MAPIM	Multi-Access PDN Connectivity and IP Flow Mobility
MBMS	Multicast/Broadcast Multimedia Service
MBMS GW	Multicast/Broadcast Multimedia Service Gateway
MBR	Maximum Bit Rate
MBSFN	Multicast/Broadcast Single Frequency Network
MCCH	Multicast Control Channel
MCE	Multi-Cell/Multicast Coordination Entity
MCS	Modulation and Coding Scheme
MIMO	Multiple-Input, Multiple-Output
MME	Mobility Management Entity
MNO	Mobile Network Operator
MPEG	Moving Picture Experts Group

MPLS	Multi-Protocol Label Switching
MPLS-TP	Multi-Protocol Label Switching Transport Profile
MTCH	Multicast Traffic Channel
MVNO	Mobile Virtual Network Operator
NACK	Negative Acknowledgement
NAS	Non-Access Stratum
NGN	Next-Generation Network
NoS	Network of Small Cells
OA&M	Operations, Administration, and Management
OPEX	Operational Expenditure
ORBIT	Open-Access Research Testbed for Next-Generation Wireless Networks
OS	Operating System
OTA	Over the Air
P-MME	Proxy Mobility Management Entity
P-SGW	Proxy Serving Gateway
PA	Paging Area
PCC	Policy Control and Charging
PCEF	Policy Control Enforcement Function
PCRF	Policy Control and Charging Rules Function
PDB	Packet Delay Budget
PDN-GW	Packet Data Network Gateway
PDR	Packet Delivery Ratio
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
PELR	Packet Error Loss Rate
PM	Power Management
PMCH	Physical Multicast Channel
ProSe	Proximity-Based Services
PS	Packet Switched
PSM	Proximity Services Management
PSS	Packet-switched Streaming Service
QAM	Quadrature Amplitude Modulation

QCI	Quality of Service Class Identifier
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RANaaS	RAN as a Service
RAT	Radio Access Technology
RAT	Remote Access Tunnel
RAU	Radio Access Unit
RC	Routing Controller
RF	Radio Frequency
RFC	Request For Comments
RRC	Radio Resource Control
RRIM	Radio Resource and Interference Management
RSS	Received Signal Strength
S-GW	Serving Gateway
S-TMSI	UE SAE Temporary Mobile Subscriber Identity
S1-AP	S1 Application Protocol
SAE	System Architecture Evolution
SAP	Service Access Point
SD-EPC	Software-Defined Evolved Packet Core
SDN	Software-Defined Network
SeGW	Security Gateway
SIB	System Information Block
SIPTO	Selected IP Traffic Offload
SO-RRIM	Self-Optimising Radio Resource and Interference Management
SON	Self-Organising Network
SQP	Sequential Quadratic Programming
STD	Standard Deviation
TA	Tracking Area
TAI	Tracking Area Identity
TAL	Tracking Area List
TAU	Tracking Area Update
TB	Transport Block

TEID	Tunnel Endpoint ID
TFT	Traffic Flow Template
TNL	Transport Network Layer
TS	Technical Specification
UE	User Equipment
UL	Uplink
uLIPA	Universal Local IP Access
UMTS	Universal Mobile Telecommunications System
USB	Universal Serial Bus
VoIP	Voice over IP
VoLTE	Voice over LTE
WCDMA	Wideband Code Division Multiple Access
WiMAX	Worldwide Interoperability for Microwave Access
WINLAB	Wireless Information Network Laboratory
WLAN	Wireless Local Area Network
X2-AP	X2 Application Protocol
XML	Extensible Markup Language

Chapter 1

Introduction

The growth in user demand for higher mobile data rates is driving Mobile Network Operators (MNOs) and network infrastructure vendors towards the adoption of innovative solutions in areas that span from physical layer techniques (e.g., carrier aggregation, Multiple-Input Multiple-Output (MIMO) schemes, etc.) to the Radio Access Network (RAN) and the Evolved Packet Core (EPC), amongst other. According to the Cisco Visual Networking Index Forecast [1], mobile data traffic is expected to grow 11-fold from 2013 to 2018, a compound annual growth rate of 61%. In terms of aggregated traffic volume, this translates into 15.9 exabytes (10^{18} bytes) per month by 2018, i.e., the equivalent of 3,965 million DVDs each month or 43,709 million text messages per second. Globally, by 2018 mobile data traffic will be equivalent to 417x the volume of global mobile traffic in 2008.

In terms of network capacity, out of a millionfold increase since 1957, the use of wider spectrum (25x increase), the division of spectrum into smaller resources (5x), and the introduction of advanced modulation and coding schemes (5x) have played a less significant role than the improvements in system capacity due to cell size reduction (1600x) [2]. This justifies the research and industrial interest in short-range, low-power cellular base stations, such as small cells. The Small Cell Forum [3] (a not-for-profit organisation aimed at accelerating the adoption of small cells in the cellular industry) defines a small cell as *a low-power wireless access point that operates in licensed spectrum, is operator-managed, and features edge-based intelligence*. In fact, the collective term *small cells* encompasses several of such network devices, namely femtocells, picocells, metrocells, and microcells – broadly increasing in size from femtocells (the smallest) to microcells (the largest). Such a fundamental shift from traditional macrocell-based networks to more heterogeneous deployments raises the need for new architectural and procedural solutions capable of providing a seamless integration of small cells into the existing cellular network infrastructure.

A novel approach for the deployment of cellular networks, known as large-scale Networks of Small Cells (NoS), has been studied in the context of the European project BeFEMTO [4]; an integrated research project funded by the European Commission with a clear industry orientation[1]. In brief, a NoS is a group of small cells that form a partially autonomous network under the administration of a Local Network Operator (LNO) that may differ from the MNO (e.g., a local authority, a corporate IT department, etc.). NoS are conceived as a complementary solution to existing cellular deployments aimed at improving network coverage, boosting system capacity, offloading traffic from the RAN and the EPC, and providing added-value services to end users. In these deployments, small cells usually feature Self-Organising Network (SON) capabilities, hence collaborating with each other to optimise the global operation of the network in terms of energy efficiency, radio resource management, control- and user-plane traffic routing, self-configuration, self-healing, etc. In this Ph.D. Thesis, we make a case for large-scale, *all-wireless* networks of small cells, where connectivity amongst small cells is provided via a wireless multi-hop backhaul. Examples of use cases for these network scenarios comprise dense urban deployments, shopping malls, corporate premises, convention centres, airports, train stations, theme parks, emergency-response scenarios, university campuses, sports venues, etc.

The research work presented in this Ph.D. Thesis dissertation has been carried out in the framework of the European project BeFEMTO. In particular, we focus on providing a substantiated answer to the following research question: ***What is the architectural and procedural framework needed to support efficient traffic and mobility management mechanisms in massive deployments of all-wireless 3GPP Long-Term Evolution (LTE) networks of small cells?*** In order to do so, we break down this research question into three key research challenges. First, we introduce a 3GPP-compliant network architecture capable of supporting large-scale, all-wireless NoS deployments in the Evolved Packet System (EPS). Essentially, this involves delegating core EPC functions onto new functional entities in the network of small

---

[1] Some of the research partners in the BeFEMTO consortium comprised Sagem Communications, NEC Europe, Telefónica R&D, DoCoMo Communications Labs Europe, T-Mobile, and Qualcomm CDMA Technologies, amongst other.

cells, as well as adapting Transport Network Layer (TNL) functionalities to the characteristics of a NoS in order to support key cellular services. This architecture, whenever possible, minimises the impact on existing 3GPP protocol stacks, functional entities, and network procedures. Secondly, we address the issue of local location management (i.e., determining the approximate location of a mobile terminal in the NoS upon arrival of an incoming connection from the core network). This involves the design, implementation, and evaluation of efficient paging and Tracking Area Update (TAU) procedures to keep track of mobile terminals in the complex scenario of an all-wireless NoS. In these schemes, efficiency is generally measured in terms of signalling traffic throughout the local NoS domain (i.e., the wireless multi-hop backhaul), as well as towards the MNO's core network. Finally, we deal with the issue of efficient traffic management in large-scale networks of small cells. In particular, we make a case for the support of direct unicast communication between LTE terminals camped on the same NoS with minimal involvement from functional entities in the core network. In addition, we propose a set of extensions to the standard 3GPP Multicast/Broadcast Multimedia Service (MBMS) in order to improve the quality of experience (QoE) of multicast traffic services in high user-density environments, such as sports venues, transportation hubs, or shopping malls. These scenarios are particularly suitable for the deployment of NoS, as mobile network operators can benefit from significant improvements in network coverage and capacity without incurring into large capital and operational expenditures (CAPEX, OPEX).

Research results have been organised in three chapters, namely *An Architectural Framework for Networks of Small Cells* (Chapter 5), *Location Management Mechanisms in Networks of Small Cells* (Chapter 6), and *Traffic Management Mechanisms in Networks of Small Cells* (Chapter 7). Each chapter provides a brief overview of the specific research problem, followed by a detailed description of the proposed solution and its corresponding analytical, simulation, or experimental performance evaluation. References to published contributions in international conferences, specialised journals, book chapters, and project deliverables have also been included at the end of each chapter. Finally, some concluding remarks and future work considerations are discussed in Chapter 8.

Chapter 2

Background

This chapter provides an overview of the concepts, mechanisms, and technologies that will be discussed in this Ph.D. Thesis dissertation. In particular, we describe the 3GPP Evolved Packet System network and protocol stack architecture (for both user and control planes), the network of small cells paradigm, and the EPS mobility and traffic management functions for unicast and multicast data services.

2.1 The 3GPP Evolved Packet System Architecture

The 3GPP Evolved Packet System is a collective term that encompasses both the LTE Radio Access Network (RAN) and the Evolved Packet Core (EPC), or core network. Occasionally, the Evolved Packet System is also referred to as System Architecture Evolution (SAE). In this section we describe the EPS network architecture, i.e., the set of functional entities that conform a full-fledged LTE mobile communications system: from the mobile terminal to the core network entities. In addition, we also discuss some of the most relevant control- and user-plane protocol stacks that provide communication services between EPS functional entities. For a detailed description of the 3GPP Evolved Packet System architecture the reader is referred to [5].

2.1.1 The EPS Network Architecture

The 3GPP Evolved Packet System is divided in two major network domains, namely the Radio Access Network and the Evolved Packet Core. Formally, the RAN encompasses the mobile terminal (or User Equipment (UE), in 3GPP terminology) and the base stations, whether macrocells (eNodeBs) or small cells, (Home eNodeBs, or HeNBs in short). In the Evolved Packet System, the RAN is also referred to as E-UTRAN (Evolved UMTS Terrestrial Radio Access Network). Note that the term *evolved UMTS* reinforces the nature of the LTE RAN as a natural evolution of the previous generation of mobile communications, namely UMTS (Universal Mobile Telecommunications System). On the other hand, the Evolved Packet Core comprises all functional entities located in the mobile operator's core network.

Figure 1 shows the 3GPP Evolved Packet System architecture, as defined in [5]. Boxes in the figure represent functional entities, i.e., a logical set of protocols and procedures that provide essential network capabilities in the RAN or the EPC. Functional entities can be implemented in a standalone manner or, alternatively, collocated with other network elements. Clouds and ovals represent networks and services (e.g., the Internet, the IP Multimedia System (IMS), internal mobile operator services, etc.). Finally, solid lines connecting functional entities to networks and services identify the standard 3GPP reference points or interfaces, as described later in the section. A 3GPP interface determines a protocol stack that adjacent EPS functional entities must implement in order to support user- and control-plane communication services between them.

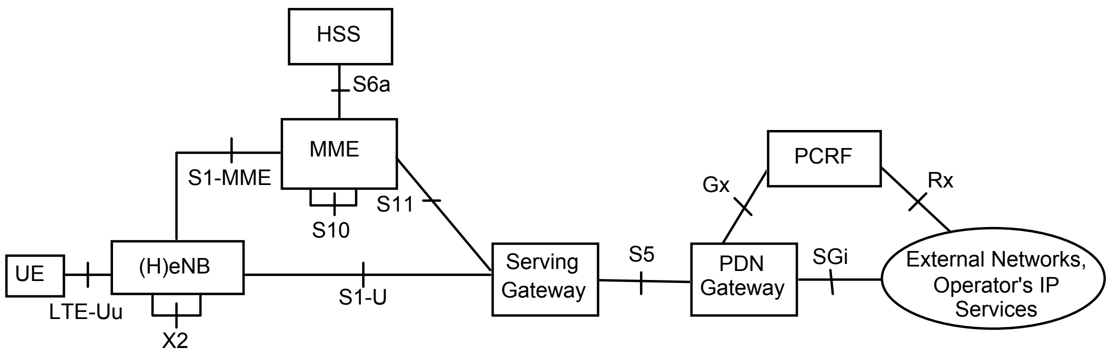


Figure 1.- The 3GPP Evolved Packet System network architecture

As seen in the figure, the main functional entities in the EPS architecture are:

- **User Equipment (UE):** this functional entity defines a mobile device capable of establishing a connection with an LTE and/or UMTS network. Although the term UE is traditionally associated with mobile phones only, it can also refer to many other mobile devices, such as tablets, laptops, LTE/UMTS USB dongles, infotainment systems, wearable devices, smart meters, etc.

- **(H)eNB:** in 3GPP terminology, the term *evolved NodeB* (or eNB, in short) refers to an LTE base station. This functional entity in the EPS architecture is responsible for various radio access network functions, such as LTE air interface provisioning, radio resource scheduling, interference management, and RAN QoS enforcement (e.g., latency, max/min bandwidth requirements for different traffic classes, interference management, etc.), amongst other.

Analogously, a *Home evolved NodeB* (HeNB, in short) is a user-deployed, low-power LTE base station with a backhaul connection to the core network aimed at providing cellular coverage in a localised area, such as an apartment or a small office. HeNBs are also referred to as small cells.

- **Mobility Management Entity (MME):** the MME is the EPS functional entity in charge of managing all control-plane traffic between the UE and the EPC. It provides a vast number of signalling services, such as session management (establishment, maintenance, and release of control- and user-plane bearers), mobility management, inter-working with legacy networks (e.g., General Packet Radio Service (GPRS) and UMTS), security, authentication, etc. The application-layer signalling protocol running between the UE and the EPC is known as the Non-Access Stratum (NAS) protocol.

- **Serving Gateway (S-GW):** the S-GW is the EPS functional entity responsible for forwarding user-plane packets from the RAN to the EPC, and vice-versa. Therefore, in the absence of traffic offloading schemes, all IP packets sent to and from a UE must traverse the S-GW before reaching their final destination. Additionally, the S-GW also acts as the local mobility anchor for data bearers during cell handovers (i.e., the transfer of user-plane data and network resources between neighbouring cells to provide session continuity during user mobility). Other relevant functions of the S-GW comprise UE context storage in idle mode (i.e., the 3GPP operational state in which the user is not engaged in a voice call or data connection), temporary buffer of downlink data during a paging procedure (i.e., the procedure by which the EPC notifies a UE of an incoming voice call or data connection), collection of user-billing information (aggregated uplink/downlink traffic data volumes), and lawful interception[2].

- **Packet Data Network Gateway (PDN-GW, or P-GW):** the PDN-GW is the EPS functional entity in charge of IP address allocation for UEs, Quality of

[2] According to the definition in [6], *lawful interception* is a mandatory network feature aimed at obtaining traffic and related information in modern telecommunications systems for the purpose of analysis or evidence in accordance with the applicable national or regional laws and technical regulations.

Service (QoS) enforcement for uplink/downlink traffic, flow-based user charging, and user-plane interfacing with the mobile operator's IP services and external networks (e.g., IMS, Internet, etc.). In addition, the PDN-GW also acts as the mobility anchor for inter-working with non-3GPP technologies, such as CDMA2000 or WiMAX. Overall, the S-GW and the PDN-GW are the two functional entities in charge of user-plane traffic management in the 3GPP Evolved Packet System.

- **Home Subscriber Server (HSS):** the HSS functional entity stores subscription data for mobile users, such as the EPS-subscribed QoS profile, access restrictions for roaming, available gateways to external networks (also referred to as *Access Point Names*, or APNs in short), etc. In some cellular network deployments, the HSS may also integrate the Authentication Centre (AuC), i.e., the EPS functional entity in charge of generating authentication vectors and security keys between mobile subscribers and the network.

- **Policy Control and Charging Rules Function (PCRF):** this EPS functional entity is responsible for policy control decision-making and flow-based charging functionalities in the Policy Control Enforcement Function (PCEF) located in the PDN-GW. Essentially, the PCRF issues authorisations to determine how the cellular network will treat user-plane traffic flows in terms of QoS levels (i.e., QoS class identifier and bit rates) according to the corresponding user's subscription profile.

As shown in Figure 1, the main reference points in the EPS architecture are:

- **LTE-Uu:** the LTE-Uu reference point defines the LTE air interface (also known as the LTE *radio* interface) between the UE and the (H)eNB.

- **X2:** the X2 reference point is located between two neighbouring (H)eNBs. This EPS interface comprises two sub-interfaces, namely the X2 user-plane (X2-U) and X2 control-plane (X2-C) in order to support various signalling and data functionalities, such as message exchange for X2-based handover assistance, inter-cell interference coordination, and IP packet forwarding.

- **S1-MME:** this is the reference point for control-plane traffic transmission and reception between the E-UTRAN and the MME.

- **S1-U:** this reference point provides user-plane tunnelling and inter-eNodeB path switching (during handover) between the E-UTRAN and the S-GW.

- **S5:** this interface provides user-plane tunnelling as well as tunnel management functions between the S-GW and the PDN-GW. The S5 reference point is also used for S-GW relocation caused by UE mobility in network deployments where the S-GW needs to connect to a non-collocated PDN-GW.

- **S6a:** this reference point between the MME and the HSS enables the transfer of user subscription data for authenticating and authorising user access to the EPS.

- **Gx:** this reference point provides the transfer of QoS policy and charging rules from the PCRF functional entity to the Policy and Charging Enforcement Function (PCEF) in the PDN-GW.

- **S10:** this reference point between MMEs provides MME relocation and MME-to-MME context information transfer services.

- **S11:** this reference point between the MME and the S-GW provides basic session and bearer management functions (creation, modification, deletion).

- **SGi:** this is the reference point between the PDN-GW and a packet data network. This packet data network may be an operator-external public/private network (e.g., Internet), or an intra-operator packet data network (e.g., IMS).

- **Rx:** this reference point is used to exchange application-level session information between the PCRF and the Application Function (AF), i.e., the network element offering application(s) that use IP bearer resources.

2.1.2 The EPS Protocol Stack Architecture

In the 3GPP Evolved Packet System architecture each reference point defines a set of logical interactions between peer functional entities (e.g., the UE and the eNodeB, the MME and the S-GW, the S-GW and the PDN-GW, etc.). These interactions are, in turn, mapped onto a set of communication protocols at different layers (physical, medium access control, network, transport, application, etc.). Essentially, this set of protocols defines the rules and procedures that EPS functional entities must implement in order to communicate with other functional entities at both control (signalling) and user-plane (data) levels. In the context of a 3GPP cellular network, the collective term *EPS protocol stack architecture* is commonly used to refer to the set of protocol stacks that provide communication between pairs of functional entities in the 3GPP Evolved Packet System architecture.

In the following subsections we describe the most relevant user- and control-plane interfaces in the 3GPP Evolved Packet System architecture, along with their corresponding protocol stacks in charge of providing communication services between peer functional entities. A full description of the EPS protocol stack can be found in [5] and [7].

2.1.2.1 User-Plane Protocol Stacks

Figure 2 shows the user-plane protocol stacks for the LTE-Uu, S1-U, S5, and SGi reference points in the 3GPP Evolved Packet System architecture. In particular, we will focus on these EPS interfaces to illustrate a conventional user-plane communication procedure between the UE and the PDN-GW.

As shown in the figure, user-plane packets exchanged between the UE and the PDN-GW are tunnelled through the various functional entities in the EPC by means of GPRS Tunnelling Protocol User Plane tunnels (GTP-U). Amongst other functions, GTP-U tunnels provide a network mechanism to enforce the corresponding QoS requirements for each traffic flow (e.g., voice, streaming video, real-time gaming, best effort data, etc.). In previous 3GPP technologies, such as GSM and UMTS, the core network has been realised through two separate domains, namely the circuit-switched (CS) domain for voice traffic and the packet-switched (PS) domain for data traffic.

Instead, LTE consolidates these two sub-domains in a single all-IP domain for combined voice and data traffic. To that extent, the entire LTE system (i.e., from the UE to the PDN-GW) is considered an end-to-end all-IP network.

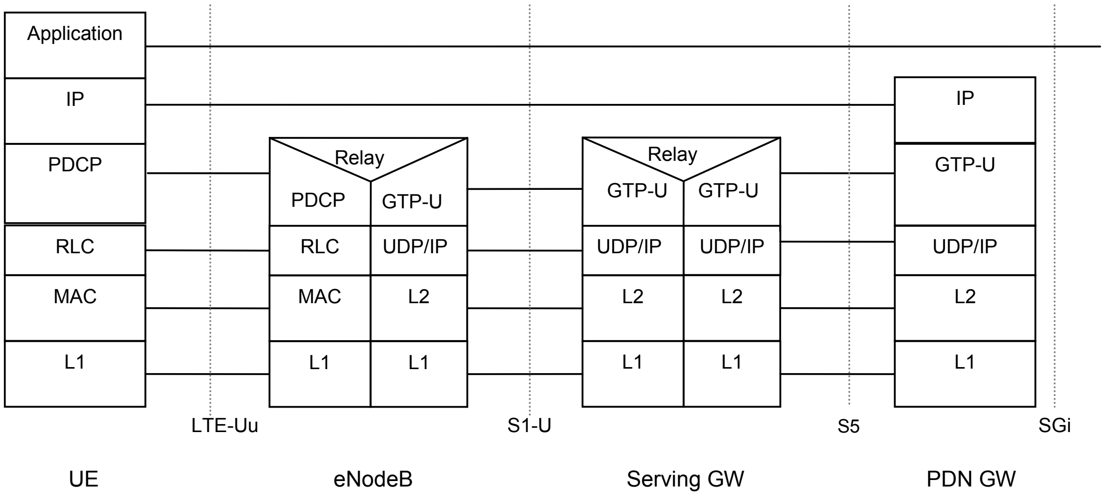


Figure 2.- The {LTE-Uu, S1-U, S5} user-plane protocol stacks

The communication protocols depicted in Figure 2 are listed below, along with a reference to their formal definition in the corresponding IETF Request for Comments (RFC) or 3GPP Technical Specification (TS):

- **Internet Protocol (IP) [8]:** in the EPS architecture, the IP protocol provides an unreliable network transport mechanism for IP packets between the UE and the PDN-GW, as well as between peer functional entities such as the eNodeB and the S-GW, and the S-GW and the PDN-GW, respectively.

- **Packet Data Convergence Protocol (PDCP) [9]:** this protocol performs header compression and decompression for user-plane data, as well as security functions (encryption, integrity), and handover support (in-sequence delivery and IP packet reordering).

- **Radio Link Control (RLC) [10]:** the RLC layer is the upper sublayer in the Layer 2 of the LTE radio protocol stack. The RLC protocol performs segmentation, concatenation, and reconstruction of PDCP PDUs. RLC also reorders RLC Protocol Data Units (PDUs) in case they have been received out of sequence due to Hybrid Automatic Repeat Request (HARQ) mechanisms in the MAC layer. With regards to Logical Link Control (LLC), an RLC protocol entity can be configured in three different modes, namely Transparent Mode (TM), Unacknowledged Mode (UM), and Acknowledged Mode (AM). For further information on these LLC modes the reader is referred to [10].

- **Medium Access Control (MAC) [11]:** the MAC layer is the lower sublayer in the Layer 2 of the LTE radio protocol stack. In brief, it constructs MAC PDUs (known as Transport Blocks (TBs)) and performs multiplexing and demultiplexing functions between logical and transport channels, i.e., its logical interfaces with the RLC and physical layers, respectively.
- **GPRS Tunnelling Protocol, User Plane (GTP-U) [12]:** the GTP-U protocol tunnels user data between eNodeBs and the S-GW, as well as between the S-GW and the PDN-GW in the Evolved Packet Core. The purpose of the GTP-U protocol is to encapsulate all IP packets in flow-specific tunnels in order to provide differentiated QoS and mobility services.
- **User Datagram Protocol (UDP) [13]:** in the EPS architecture, the UDP protocol is used to transport GTP-U data between the eNodeB and the S-GW, and between the S-GW and the PDN-GW. It provides multiplexing and data integrity (e.g., checksum) services for user-plane traffic over the S1-U and S5 reference points.

Another relevant interface in the 3GPP Evolved Packet System is the X2 interface. As previously described, this reference point is sub-divided in a user-plane (X2-U) and a control-plane (X2-C) interface. As far as user-plane traffic is concerned, the X2-U interface provides non-guaranteed delivery of user-plane PDUs between neighbouring (H)eNBs. This is commonly used for user-plane traffic forwarding between (H)eNBs during X2-based handovers. Figure 3 shows the protocol stack associated with the X2-U reference point. As shown in the figure, this protocol stack is essentially identical to the S1-U protocol stack described in Figure 2. To this extent, the transport network layer is also built on GTP-U over UDP/IP to carry user-plane PDUs.

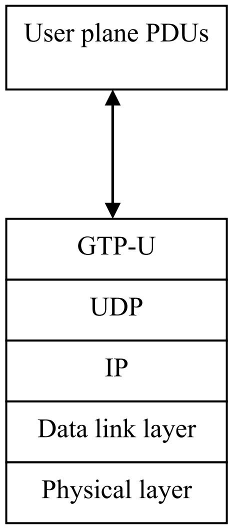


Figure 3.- The X2-U protocol stack

2.1.2.2 Control-Plane Protocol Stacks

Figure 4 shows the control-plane protocol stack architecture for the LTE-Uu and S1-MME reference points. The former implements the LTE air interface between the UE and the eNodeB, whilst the latter provides logical connectivity between the eNodeB and the MME, thus enabling key cellular services such as user registration, mobility, authentication, and session establishment. For the sake of providing a brief description of the most relevant control-plane procedures between the UE and the EPC, we will focus on the LTE-Uu and S1-MME reference points only. For a comprehensive description of all control-plane protocol stacks in the Evolved Packet System architecture the reader is referred to [5] and [7].

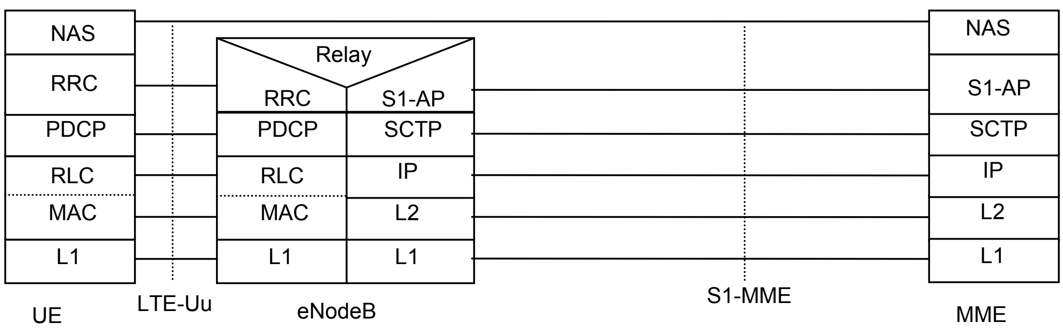


Figure 4.- The {LTE-Uu, S1-MME} control-plane protocol stack

The protocols depicted in Figure 4 are described below. A reference to their formal definition in the corresponding IETF Request for Comments (RFC) or 3GPP Technical Specification (TS) has also been provided:

- **Non-Access Stratum (NAS) [14]:** the NAS protocol is the key control-plane application protocol between the UE and the Evolved Packet Core. Amongst other functions, the NAS protocol handles network procedures such as Public Land Mobile Network (PLMN) selection, tracking area update, paging, authentication, as well as EPS bearer establishment, modification, and release.

- **Radio Resource Control (RRC) [15]:** RRC is a critical control-plane protocol in the LTE radio access network. Amongst other functions, RRC is in charge of the establishment, maintenance, and release of radio resources in the E-UTRAN (i.e., between the UE and the eNodeB), as well as of transparently relaying system information and common/dedicated NAS information between the UE and the MME. In addition, RRC provides measurement configuration and reporting services for intra-frequency, inter-frequency, and inter-RAT mobility procedures (e.g., handovers, cell redirections).

- **S1 Application Protocol (S1-AP) [16]:** the S1-AP protocol provides the necessary application-layer signalling between the E-UTRAN and the EPC to perform network functions such as bearer establishment, mobility, and high-layer signalling transport via the Non-Access Stratum protocol, amongst others.

- **Stream Control Transport Protocol (SCTP) [17]:** the SCTP transport protocol provides a reliable signalling message delivery service between the eNodeB and the MME in the Evolved Packet System network architecture. Compared with traditional connection-oriented transport protocols such as TCP, SCTP enables multistream handling in order to provide enhanced features like transport network redundancy, head-of-line blocking avoidance, and multi-homing. For a detailed description of the SCTP protocol the reader is referred to [17].

Another key control-plane interface in the 3GPP Evolved Packet System network architecture is the X2-C interface (Figure 5). This reference point provides control-plane communication services between neighbouring (H)eNBs. This is particularly useful during X2-based handovers, where (H)eNBs need to exchange control messages in order to provide session continuity between the source and target cells.

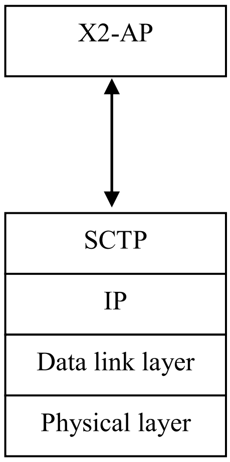


Figure 5.- The X2-C protocol stack

Essentially, the X2-C protocol stack is identical to the S1-MME protocol stack, with the exception of the S1-AP protocol being replaced with the X2-AP protocol in the application layer. This signalling protocol supports various functions, such as intra-LTE mobility support for UEs in RRC active mode, context transfer from source to target (H)eNBs during X2-based handovers, control of user-plane tunnels between source and target (H)eNBs, load management, inter-cell interference coordination information, etc. A detailed description of the X2-AP protocol can be found in [18]. Analogously to the case of the S1-MME protocol stack, the transport network is also built on SCTP on top of IP.

2.2 The Small Cells Paradigm

Mobile network operators have recently deployed a large number of standalone small cells to target indoor residential/small enterprise scenarios where network coverage was either poor or non-existent. In fact, in November 2012, the Small Cell Forum reported that the total number of deployed standalone small cells had exceeded the number of traditional mobile base stations worldwide [19]. Although standalone small cell deployments can help to improve cellular coverage and capacity in very localised areas (e.g., apartments, small offices), they do not fully exploit the potential of applying small cell technology to large-scale scenarios.

2.2.1 Standalone Small Cells

Standalone small cells were first introduced in Release 8 of 3GPP Technical Specifications. Initially, standardisation bodies and industry forums like 3GPP and the

Small Cell Forum coined the term *femtocell* to refer to user-deployed, low-power base stations operating in licensed spectrum. However, they later adopted the more generic term *small cell* to target not only standalone residential/small office scenarios, but also larger enterprise and metropolitan deployments. For the sake of simplicity, in this Ph.D. Thesis dissertation we will use the terms *small cell* and *HeNB* indistinctively.

One of the most challenging aspects of standalone small cells in Release 8 of 3GPP Technical Specifications was the design of an architectural framework to support the integration of ubiquitous residential small cell deployments in the Evolved Packet System. As previously mentioned, standalone small cells are user-deployed network elements operating in licensed spectrum with a backhaul connection (e.g., ADSL, cable, Ethernet, etc.) to provide connectivity with functional entities in the Evolved Packet Core. To this extent, the initial rollout of standalone small cell deployments posed serious challenges on various network aspects such as architecture, security, operations and management, and scalability. Figure 6 shows the E-UTRAN HeNB network architecture, as defined in Release 8 of 3GPP Technical Specifications [7].

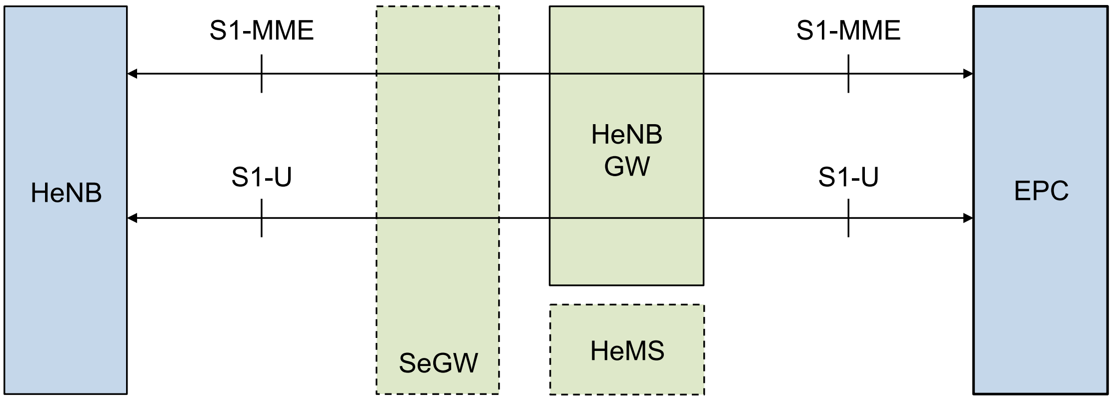


Figure 6.- The E-UTRAN HeNB network architecture

As seen in the figure, the E-UTRAN architecture may deploy a Home eNB Gateway (HeNB GW) to allow the S1 interface between the HeNB and the Evolved Packet Core to scale to support a large number of HeNBs. From the functional point of view, the HeNB GW acts as a concentration node for control-plane traffic between HeNBs and the Evolved Packet Core, in particular for the S1-MME interface. Instead, the S1-U interface might be terminated at the HeNB GW or, alternatively, a direct logical user-plane connection between the HeNB and the S-GW may be established.

When deployed, the HeNB GW acts as a transparent functional entity in the 3GPP Evolved Packet System architecture. Thus, the HeNB GW appears to the HeNB as a standard MME. Analogously, the HeNB GW appears to the MME as a regular HeNB. Some mobile network operators might opt to deploy cellular networks that combine HeNBs that are directly connected to the core network with HeNBs that connect to the EPC via a HeNB GW. In both scenarios, the user- and control-plane protocols between all EPS functional entities (i.e., the MME, the S-GW, and the HeNBs) remain untouched, as discussed later in the section. Figure 7 shows a small cell deployment scenario that combines HeNBs directly connected to the Evolved Packet Core with HeNBs connected to the core network via HeNB GWs.

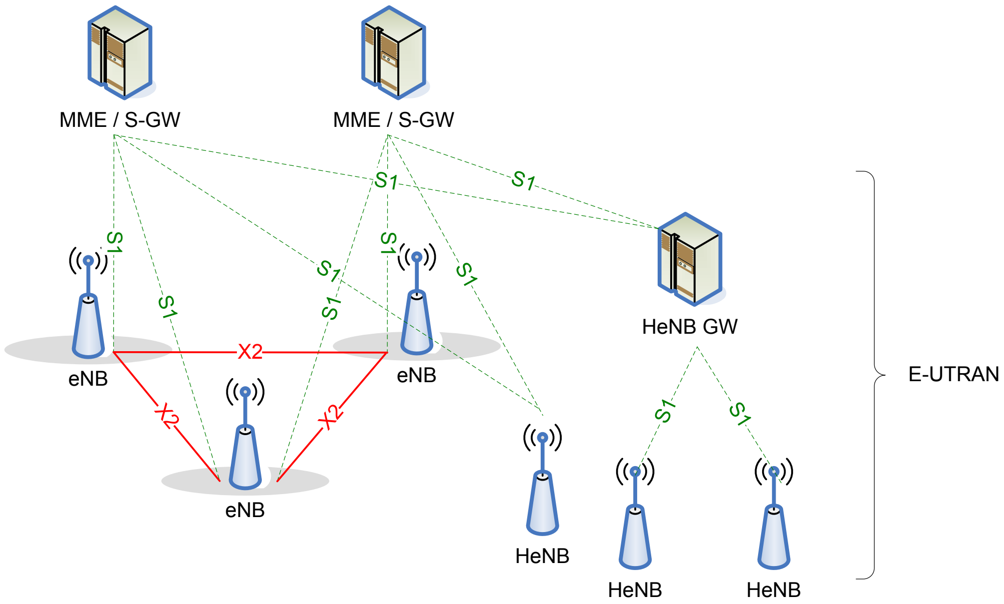


Figure 7.- Overall E-UTRAN architecture with deployed HeNB GW

In addition to the HeNB GW, Release 8 of 3GPP Technical Specifications also supports the deployment of a HeNB Management System (HeMS), a centralised functional entity to operate and maintain all HeNBs in the cellular network, as well as a Security Gateway (SeGW). This optional functional entity is located at the border of the mobile operator's core network and provides a secure tunnelling between HeNBs and the HeMS, and between HeNBs and the MME. After completing a successful mutual authentication procedure between the HeNB and the SeGW, the SeGW connects the HeNB to the operator's EPC. From this moment onwards, all connections between the HeNB and the core network are tunnelled through the SeGW functional entity.

As previously discussed, the presence of the HeNB GW is transparent to both the MME and the HeNB. In practice, this means that neither the MME's nor the HeNB's control- and user-plane protocol stacks need to be modified should the mobile network operator opt to deploy a HeNB GW. In terms of protocol stack architecture, the HeNB GW must implement the S1-U (user) and S1-MME (control) reference points, as well as their corresponding protocol stacks. Figure 8 and Figure 9 show the S1-U and S1-MME protocol stacks between the HeNB and the HeNB GW, respectively. Note that the HeNB GW may optionally terminate the user plane towards the HeNB and the S-GW, and provide a relay function for user-plane data.

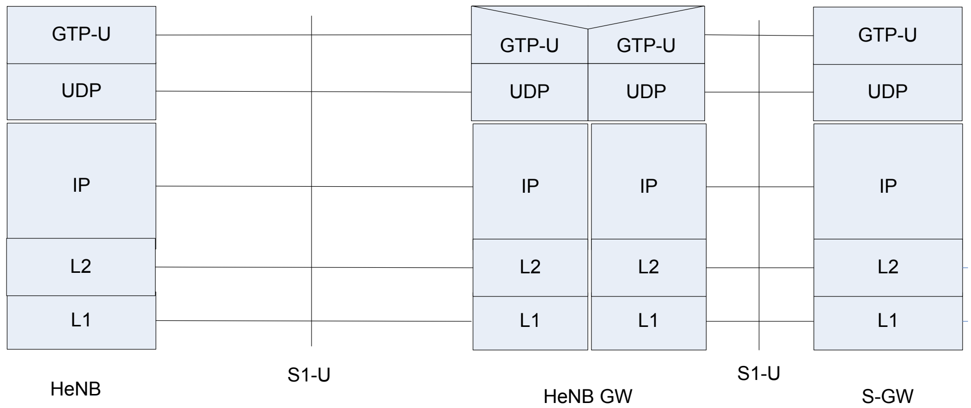


Figure 8.- The S1-U protocol stack in the {HeNB, HeNB GW, S-GW} entities

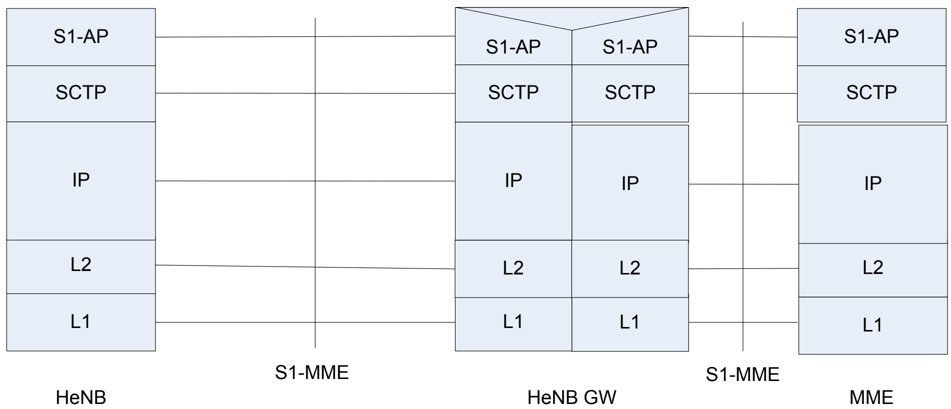


Figure 9.- The S1-MME protocol stack in the {HeNB, HeNB GW, S-GW} entities

2.2.2 Networks of Small Cells

As opposed to standalone small cells, a Network of Small Cells (or NoS, in short) is a group of small cells connected via a local wired or wireless network, and typically belonging to the same administrative domain. Small cells perform functions like radio resource and mobility management in a cooperative way and, mainly, by means of local communication (i.e., by minimising the involvement of functional entities in the Evolved Packet Core). Contrary to conventional standalone deployments, where each small cell operates in an autonomous and uncoordinated manner, networks of small cells optimise global network performance by allowing cooperation in a self-organising fashion to provide enhanced cellular services to a larger number of users. This makes them different from collections of geographically close standalone femtocells that are not coordinated or get coordinated over the MNO's core network.

Mobile network operators can leverage NoS to improve network coverage and boost capacity in high user-density areas, such as transportation hubs, corporate premises, sports venues, university campuses, dense urban scenarios, convention centres, financial districts, shopping malls, etc. For MNOs, networks of small cells are a cost-efficient and easy to deploy complementary solution to traditional macrocell networks due to more affordable RAN equipment (i.e., small cells vs. macrocells), lower access fees to urban furniture (e.g., traffic lights, lampposts, building façades, venue-specific infrastructure, etc.), and leaner network configuration, operation, and maintenance. On the other hand, larger groups of mobile users can benefit from NoS by achieving a better QoE due to their physical proximity to small cell equipment.

In a network of small cells, some of the functionalities that are traditionally under the control of the Evolved Packet Core need to be transferred to new functional entities in the NoS. This is done to reduce the volume of user data and signalling traffic reaching key EPC control- and user-plane entities, such as the MME and the S-GW. Note that, in standalone residential/enterprise deployments, the cost and scalability issues associated with the backhaul link towards the EPC may not necessarily be an issue. However, in massive deployments of NoS, this might be a serious concern due to the high volume of user- and control-plane traffic towards the EPC. Thus, one of the most challenging issues in a NoS is to define a network architecture that allows functional entities under the control of a *local network operator* (e.g., municipalities, IT departments, venue management authorities, etc.) to run MNO-controlled procedures, whilst complying with 3GPP Technical Specifications. Further challenges comprise (amongst other) user tracking and paging in the NoS domain (local location management), session continuity during handoffs between small cells (handover management), access to local services (Local IP Access, LIPA), local traffic exchange between UEs camped on the same NoS (traffic management), and scalable Multicast/Broadcast Multimedia Services (MBMS).

Figure 10 shows the network architecture of an all-wireless network of small cells in the context of the 3GPP Evolved Packet System. The core network comprises all standard functional entities (e.g., MME, S-GW, P-GW), as well as some of the most relevant 3GPP interfaces (S1-MME, S1-U, S5, etc.). In particular, communication between the Evolved Packet Core and the NoS is provided by the S1-MME (control plane) and S1-U (user plane) logical interfaces, which must traverse an Internet Service Provider (ISP) network. Chapter 5 provides a detailed description of the functional entities and reference points in a network of small cells. For a comprehensive description of the functional entities in the 3GPP Evolved Packet System the reader is referred to [5].

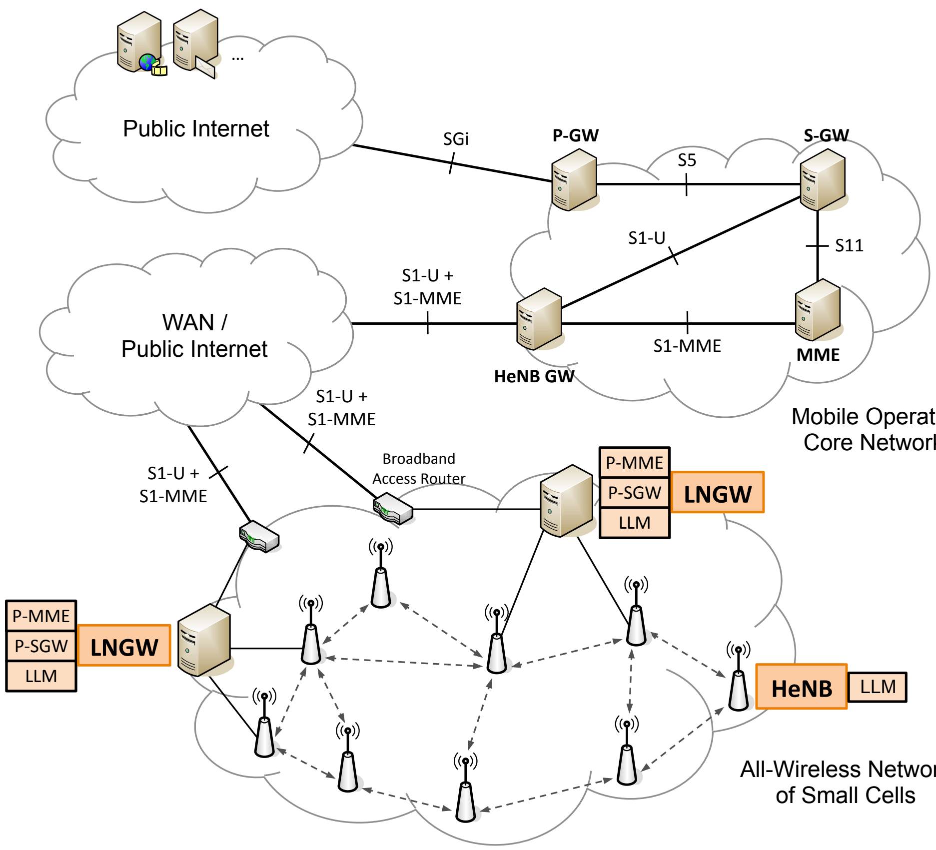


Figure 10.- Network architecture of an all-wireless network of small cells

In a NoS, a new functional entity called the *Local Network of Small Cells Gateway* (LNGW) has been introduced to act as an interface between the NoS and the EPC. This is explained in detail in Chapter 5. Furthermore, if the NoS is all-wireless, small cells are connected to each other without the need for cable laying, hence forwarding data packets towards/from LNGWs over a wireless multi-hop backhaul. These generalised all-wireless NoS deployments allow multi-radio small cells to be fully exploited. In recent years, it has become a common practice for network infrastructure vendors to implement small cells that, in addition to the 3GPP radio access technology modem (e.g., LTE or WCDMA), they also deploy a non-3GPP wireless modem [19], [21]. This additional wireless modem, whether conventional Wi-Fi (802.11a/b/g/n/ac), carrier-grade Wi-Fi, or any other radio access technology, can be leveraged not only to provide wireless access to end users, but also –or alternatively– to deploy a local wireless backhaul amongst neighbouring small cells.

Compared with conventional standalone small cells scenarios, NoS allow more cost-effective and flexible deployments by building a network between nearby small cells, hence allowing multiple HeNBs to share the same connection to the mobile operator's core network [22]. Essentially, this introduces several advantages. First, it reduces capital (CAPEX) and operational (OPEX) expenditures for both the MNO and the local network operator. Second, it offloads signalling traffic from the core network to the network of small cells. Third, it helps to offload data traffic from the MNO's core network when combined with local IP access (LIPA) and Selected IP Traffic Offload (SIPTO) [23]. In addition, if the backhaul links between neighbouring small cells are wireless, the resulting NoS is also heterogeneous (in a generalised way) in the sense that it combines 3GPP and non-3GPP radio access technologies to build a 3GPP-oriented wireless mesh network. This way, operators can eliminate the need of cable laying in order to reach each HeNB in the network of small cells.

2.3 Mobility Management in the Evolved Packet System

Mobility management is a key function in cellular networks. Traditionally, this collective term encompasses both *location management* (i.e., tracking the approximate location of a UE in a cellular network) and *handover management* (i.e., the set of procedures aimed at providing seamless voice/data session continuity when a user moves between neighbouring base stations). As shown in Figure 11, location and handover management functions are tightly coupled to the main two RRC modes in which a user equipment can be found when camped on a cellular network: *idle* and *active* [24].

A UE is in RRC idle mode (or *idle* mode, in short) when there are no dedicated radio resources established with the network [24]. A UE in idle mode is able to listen to and decode all broadcast channels, derive system-wide broadcast information, and initiate/receive calls towards/from the network. On the other hand, a UE is in RRC active mode (or *active* mode, in short) if it has established dedicated radio resources with the network [24]. Usually, these resources have been previously established as a result of an ongoing voice call or data connection (e.g., web browsing, videoconference, file upload/download, voice call, etc.). As a general rule, location management procedures are executed in idle mode, whilst handover management procedures take place when the UE is in active mode.

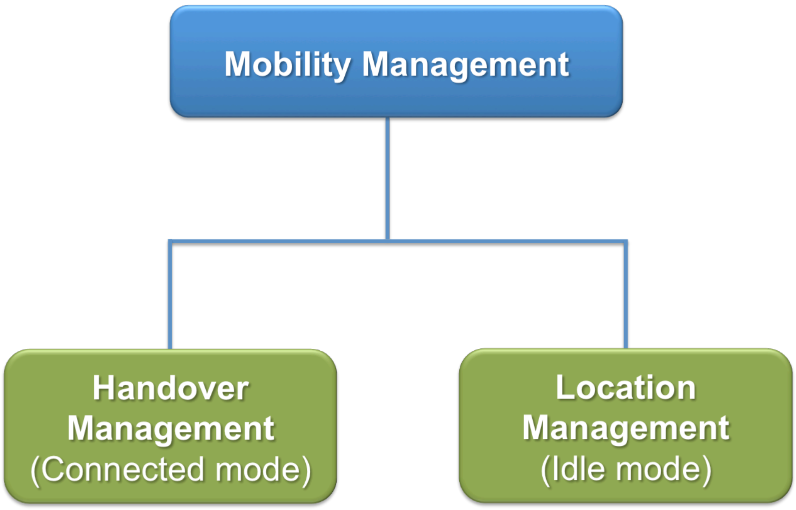


Figure 11.- Mobility management functions vs. RRC states

2.3.1 Tracking Area Update

The two key location management procedures in the 3GGP Evolved Packet System are *Tracking Area Update* (TAU) and *paging*. Both procedures are tightly coupled in the sense that tracking area update provides the mobile network operator with UE location information that will be later utilised by the paging procedure.

The goal of the TAU procedure is to report the approximate location of the UE to the MME functional entity in the Evolved Packet Core. This is done to enable the bearer establishment procedure when a voice call or data connection needs to be initiated with the UE. The TAU procedure is triggered by the UE every time the mobile terminal enters a new Tracking Area (TA), or cluster of neighbouring cells. The size, number, and shape of these tracking areas are normally static and well defined by the mobile network operator. When a connection between the core network and a UE needs to be established, the MME functional entity does not need to know the exact location of the UE on a per-cell basis, but only on a TA basis instead. This is done to prevent UEs from performing too many TAU procedures due to user mobility, thus preserving battery life and reducing signalling traffic throughout the radio access and core networks. Once the approximate location of a UE in the cellular network has been determined (i.e., on a per-TA basis), the Mobility Management Entity in the EPC is ready to initiate the paging procedure. Figure 12 describes the sequence of NAS messages in a standard 3GPP Tracking Area Update procedure. For a detailed description of the TAU procedure (including potential S-GW and MME relocations) the reader is referred to [5].

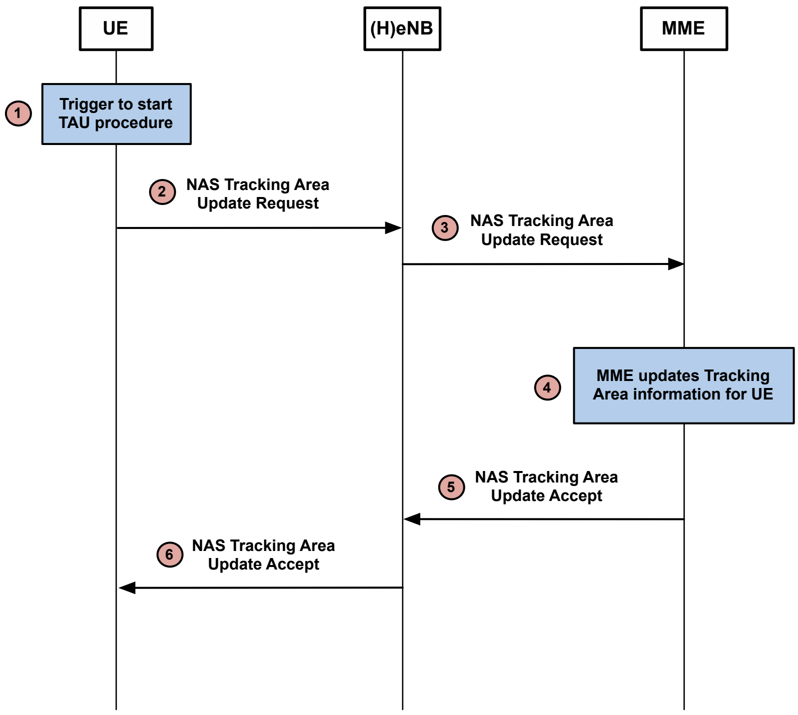


Figure 12.- The 3GPP Tracking Area Update procedure

Each step in the Tracking Area Update procedure is described below:

- **Step 1:** the UE determines if a TAU procedure needs to be initiated. Although there are multiple conditions that can trigger a TAU procedure (as defined in [5]), the most common one is to check whether the UE has entered a new tracking area. The UE can do so by obtaining the Tracking Area Identity (TAI) of the serving cell from the System Information Blocks (SIBs) transmitted over the Physical Downlink Shared Channel (PDSCH) and comparing it with the last stored TAI value. If these two values differ, the UE concludes that it has entered a new tracking area and, thus, a TAU procedure must be initiated.

- **Step 2:** the UE sends a *NAS Tracking Area Update Request* message to the MME via the (H)eNB. As described in Section 2.1.2, NAS messages exchanged between the UE and the (H)eNB are sent encapsulated in RRC messages. Essentially, the purpose of the *NAS Tracking Area Update Request* message is to notify the MME that it needs to update the TAI entry of the corresponding UE in its location management database.

- **Step 3:** the (H)eNB transparently relays the *NAS Tracking Area Update Request* message to the MME functional entity in the core network.

- **Step 4:** upon reception of the *NAS Tracking Area Update Request* message, the MME updates the TAI entry of the corresponding UE in its location management database. From this moment onwards, the UE is effectively registered to the new tracking area.

- **Step 5:** the MME in the core network sends a *NAS Tracking Area Update Accept* message to the UE via the (H)eNB. Amongst other data, this message contains the necessary Information Elements (IE) to acknowledge the successful registration of the originating UE into the new tracking area.

- **Step 6:** the (H)eNB relays the NAS message to the originating UE encapsulated in an RRC message. Upon reception of this message, the UE updates its local last visited Tracking Area Identity. Effectively, this completes the standard 3GPP Tracking Area Update procedure.

2.3.2 Paging

The goal of the 3GPP paging procedure is to notify a UE of an incoming voice call or data connection. After successful completion of this procedure, the destination UE initiates the establishment of the corresponding dedicated EPC and radio resources between the UE and the core network in order to support the incoming voice/data services. Since the location of the UE within the cellular network is known on a tracking area basis (as described in the previous section), the MME needs to send an indication message (i.e., a *paging* message) to each cell in the UE's last registered TA informing about the incoming voice or data connection. Essentially, when a (H)eNB receives a paging message from the MME, it sends a broadcast paging notification over the LTE-Uu interface containing the UE identity. Once the broadcast paging notification reaches the destination UE, the terminal establishes the necessary dedicated radio and EPC resources with the network, thus effectively moving from idle to active state. Other UEs in the destination tracking area (i.e., non-destination UEs) simply ignore the broadcast paging notification. Figure 13 shows the message sequence chart describing the 3GPP standard paging procedure, as defined in [7].

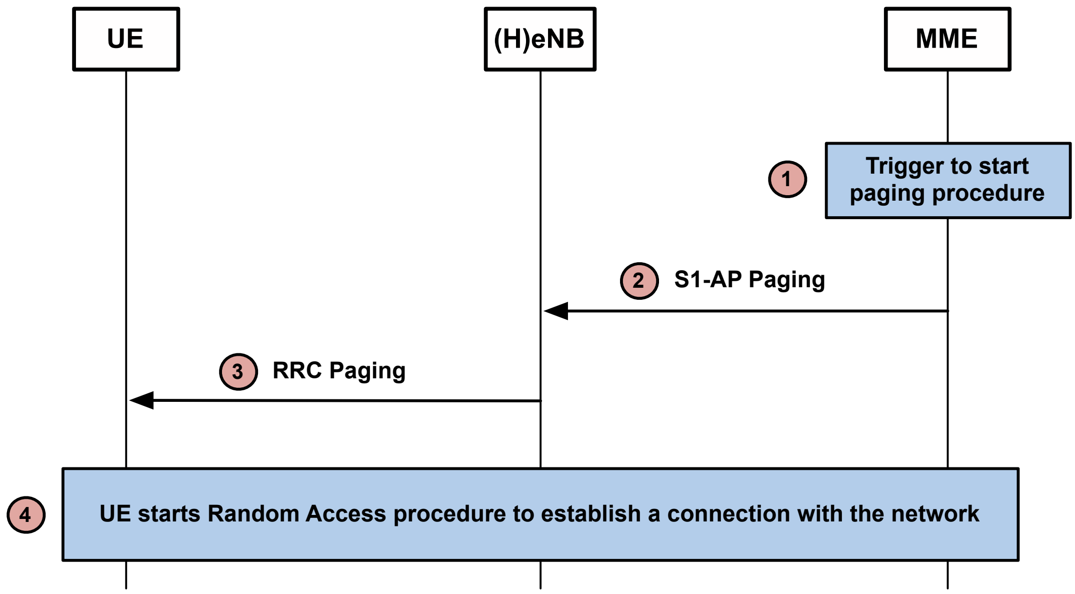


Figure 13.- The 3GPP paging procedure

The different steps in the paging procedure are described below:

- **Step 1:** the MME determines if a destination UE must be paged. This condition is verified when the MME receives a *GTP-C Downlink Data*

Notification message from the S-GW functional entity, i.e., when the core network starts sending IP packets to the destination UE.

- **Step 2:** the MME processes the *GTP-C Downlink Data Notification* message from the S-GW and sends an *S1-AP Paging* message to each (H)eNB with cells belonging to the tracking area to which the UE is currently registered. Although each (H)eNB can contain cells that belong to different tracking areas, each cell can only belong to one TA.

- **Step 3:** the (H)eNB receives the *S1-AP Paging* message from the MME and sends an *RRC Paging* message to the destination UE(s) over the LTE-Uu interface. Note that the *RRC Paging* message may contain multiple paging records in order to page multiple UEs.

- **Step 4:** the UE receives and decodes the *RRC Paging* message from the (H)eNB (this message is sent over the PDSCH). If the UE does not find its own identity in the paging records contained in the *RRC Paging* message it returns to idle mode. Otherwise, it triggers a Random Access procedure in order to establish an RRC connection with the Evolved Packet Core, thus effectively switching from idle to active mode.

2.3.3 Handover

The most critical mobility management procedure in active mode is the handover (HO) procedure. Handovers (also known as *handoffs*) ensure that mobile terminals in RRC active mode are always physically connected to the best serving cell in terms of QoS metrics such as received signal strength or received signal quality. Furthermore, the handover procedure implements the necessary steps to seamlessly transfer all ongoing voice calls and data connections from the source to the target (H)eNB whenever user mobility requires so. The *source* (H)eNB is the cell to which the mobile terminal is currently connected. Instead, the *target* (H)eNB is the cell to which the mobile terminal will be connected once the handover procedure is successfully completed.

From the point of view of resource management (at both RAN and EPC levels), the handover procedure allocates the necessary radio resources in the target cell (*HO* *preparation*), notifies the source cell about the handover operation (*HO execution*), and switches all traffic paths from the source to the target cell in the core network (*HO completion*). From the architectural point of view, the handover procedure involves multiple functional entities in the radio access and core networks, namely the UE, the source and target (H)eNBs, and the (source and target) MMEs and S-GWs, in case of MME and/or S-GW relocation. Figure 14 shows the message sequence chart of an intra-MME/S-GW handover, i.e., a handover in which there are no MME or S-GW relocations. This message sequence chart assumes that a logical X2 interface has previously been established between the source and target (H)eNBs. For a detailed description of the LTE handover procedure in various MME and S-GW relocation scenarios the reader is referred to [7]. The steps involved in an intra-MME/S-GW handover are discussed below:

- **Step 1:** the source (H)eNB configures the UE measurement reporting procedures that will be used to assist the UE connection mobility function. The aim of the measurement reports is to inform the source (H)eNB about the received signal strength/quality of the serving and neighbouring (H)eNBs, whether intra- and inter-frequency. Measurement reports can be configured in event-triggered and periodic modes.

- **Step 2:** the UE generates a measurement report (event-triggered or periodic) in order to inform the source (H)eNB about the received signal strength/quality of the serving and/or inter/intra-frequency neighbouring cells.

- **Step 3:** based on the received measurement reports, the source (H)eNB determines if a handover procedure to a more suitable cell must be initiated.

- **Step 4:** the source (H)eNB issues a *Handover Request* message to the target (H)eNB containing the necessary information to prepare the handover in the target cell. The aim of this message is to instruct the target (H)eNB to perform admission control functions in order to determine if it can grant radio resources to the new user.

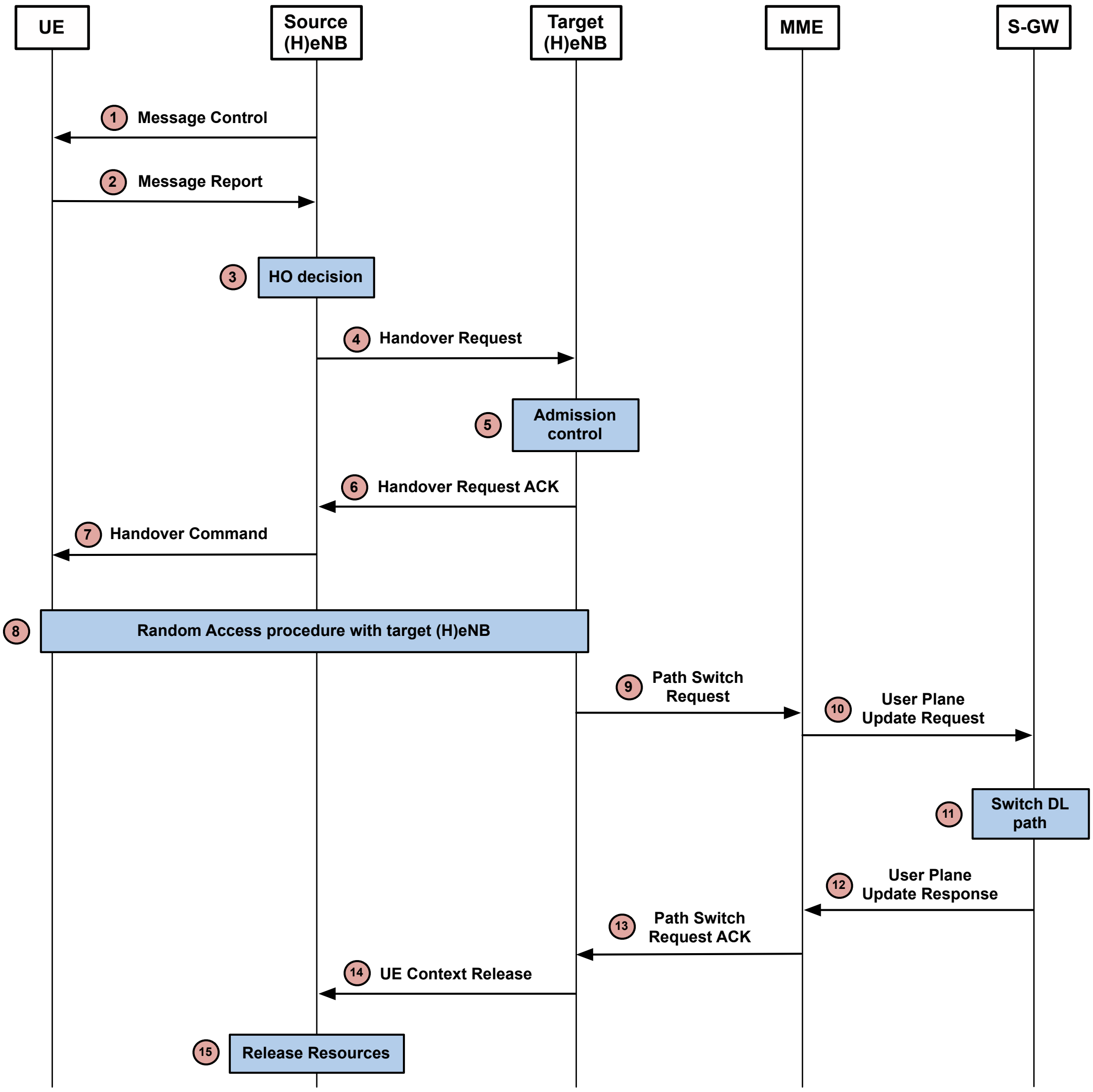


Figure 14.- Intra-MME/S-GW handover message sequence chart

- **Step 5:** the target (H)eNB performs admission control functions based on its current available resources and the information provided by the source (H)eNB. If resources can be granted, the UE and the cellular network proceed with the handover. Otherwise, the handover is cancelled.

- **Step 6:** the target cell configures its L1/L2 resources accordingly and sends the *Handover Request Acknowledge* message to the source (H)eNB. In addition to the necessary configuration parameters, this message also encapsulates a transparent container to be forwarded to the UE in the form of an RRC message in order to perform the handover (step 7).

- **Step 7:** the source (H)eNB decapsulates the transparent container received from the target (H)eNB and forwards the handover command message (*RRC*

Connection Reconfiguration) to the UE. Upon reception of this message, the UE detaches from the old cell (source (H)eNB) and attempts to synchronise with the new cell (target (H)eNB).

- **Step 8:** the UE initiates a Random Access procedure to connect to the target (H)eNB. In addition, the UE derives the target cell-specific keys and configures the selected security algorithms to be used in the new connection.

- **Step 9:** the target (H)eNB sends a *Path Switch* message to the MME to inform that the UE has changed its serving cell. The aim of this message is to trigger the necessary procedures in the Evolved Packet Core to enable user-plane traffic redirection from the S-GW to the new serving cell, and vice-versa.

- **Step 10:** the MME sends a *User Plane Update Request* message to the S-GW in order to enable user-plane traffic redirection.

- **Step 11:** the S-GW switches the downlink data path to the target cell. In addition, the S-GW sends one or more *end marker* packets on the old path to the source (H)eNB in order to indicate the release of any associated radio and network resources.

- **Step 12:** the S-GW sends an *Update User Plane Response* message to the MME, thus acknowledging the change of user-plane data path towards the new cell.

- **Step 13:** the MME sends a *Path Switch Request Acknowledge* message to the target (H)eNB to confirm the change of user-plane data path towards the new cell.

- **Step 14:** by sending the *UE Context Release* message, the target cell informs the source (H)eNB about the success of the handover procedure. This message triggers the release of any unused resources in the source (H)eNB.

- **Step 15:** upon reception of the *UE Context Release* message, the source (H)eNB releases all radio and control-plane related resources associated with the UE context.

2.4 Traffic Management in the Evolved Packet System

The term *traffic management* refers to the architecture, protocols, and network procedures needed to support user-plane unicast and multicast/broadcast services in the context of the 3GPP Evolved Packet System. First, we will focus *unicast traffic management*, with a particular emphasis on the EPS bearer service (i.e., the standard QoS-enforcing mechanism in the Evolved Packet System). Secondly, we will discuss the Multicast/Broadcast Multimedia Service (MBMS), the standard 3GPP mechanism to support *multicast* and *broadcast traffic management* in the Evolved Packet System.

2.4.1 Unicast Traffic Management

In communication networks, *unicast* traffic is traditionally exchanged between a pair of single nodes, i.e., the source and destination nodes. In the 3GPP Evolved Packet System, IP traffic exchange between the UE and the P-GW as a result of a voice call or data connection is, by definition, unicast. However, the EPS also supports *multicast/broadcast* services, where a larger number of UEs simultaneously receive data from the same server (e.g., live streaming content, audio/video broadcast, news reel, etc.). In this subsection we introduce the architectural and procedural framework needed to establish unicast data connections between a UE and the P-GW, as well as the standard QoS-enforcing mechanism needed to provide quality of service across functional entities in the Evolved Packet System.

2.4.1.1 The EPS Bearer Service

In cellular networks (and, in general, data networks as a whole), QoS mechanisms define traffic priorities for customers and/or services during times of high congestion. These mechanisms apply to both premium customers who are willing to pay more in order to benefit from a better quality of experience in their mobile devices, as well as to critical communication services that require higher traffic priority across the network (e.g., voice over IP (VoIP), video calls, real-time gaming, etc.). For instance, if a user is engaged in a voice call while browsing the web, the mobile network operator may treat each data flow differently in order to meet the specific QoS demands of each service. This is achieved by applying the corresponding QoS-enforcing mechanisms across all adjacent functional entities in the user-plane data path (e.g., UE, eNB, S-GW, and P-GW). In the context of the 3GPP Evolved Packet

System, these QoS-enforcing mechanisms are collectively known as *the EPS bearer service* [7].

In 3GPP terminology, a *bearer* is a network abstraction used to represent an IP user-plane data flow associated with a specific set of QoS requirements. When a UE and a peer entity (e.g., another UE, an internal server in the MNO's network, a host on the Internet, etc.) initiate a voice call or data connection, functional entities in the EPS trigger the establishment of bearers or *virtual pipes* across the radio and core networks in order to enforce the corresponding QoS policies for all data packets transmitted over that connection. In practice, the scope of applicability of the EPS bearer service is limited to the user-plane functional entities between the UE and the P-GW. Consequently, the EPS bearer service does not support QoS enforcement on functional entities that lie beyond the scope of the MNO network (e.g., a web server on the Internet). Figure 15 shows the EPS bearer service architecture, as defined in [7].

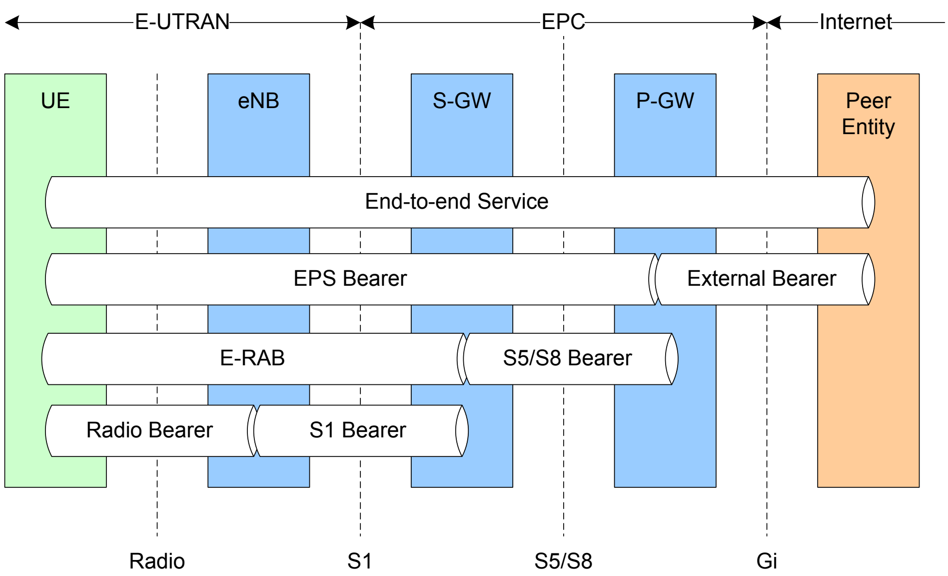


Figure 15.- The EPS bearer service architecture

The bearers depicted in Figure 15 are defined as follows:

- **Radio bearer:** a radio bearer carries user-plane data between the UE and the (H)eNB via the LTE-Uu interface. In practice, radio bearers enforce the required QoS policy by appropriately configuring the RLC, MAC, and PDCP layers in the (H)eNB and UE protocol stacks (e.g., radio resource allocation, scheduling policy, ARQ mechanism, etc.).

- **S1 bearer:** the S1 bearer carries data between the (H)eNB and the S-GW over the S1-U interface. Essentially, S1 and S5 bearers (i.e., non E-UTRAN bearers) enforce the required QoS policies by triggering the necessary QoS mechanisms in the underlying transport network (e.g., DiffServ, Multi-Protocol Label Switching (MPLS), etc.). Since these are transport network layer mechanisms, the MNO must ensure that the underlying transport network is capable of providing the necessary QoS guarantees to upper layers.
- **Evolved Radio Access Bearer (E-RAB):** the E-RAB is the result of concatenating the radio and S1 bearers. Thus, the E-RAB carries user-plane data between the UE and the S-GW according to the QoS policy defined by the combination of the S1 and radio bearers.
- **S5 bearer:** the S5 bearer carries data between the S-GW and the P-GW functional entities over the S5 interface. If the S-GW and P-GW are collocated in a standalone network element, the S5 bearer is absent. Analogously to the S1 bearer, the S5 bearer enforces the corresponding QoS policies by triggering the appropriate QoS mechanisms in the underlying transport network.
- **EPS bearer:** the EPS bearer is the fundamental QoS-enforcing building block in the 3GPP EPS bearer service. Analogously to the E-RAB, the EPS bearer is the result of concatenating the radio, S1, and S5 bearers. In terms of QoS, the EPS bearer carries data between the UE and the P-GW according to the QoS policy defined by the combination of the S5, S1, and radio bearers.
- **External bearer:** the external bearer carries data between the P-GW and the corresponding peer entity in an external network (e.g., the Internet). Since the scope of the external bearer falls beyond the boundaries of the 3GPP Evolved Packet System we will not be consider it further in this section.

Additionally, bearers can be classified into two categories based on the nature of the quality of service that they provide, namely:

- **Minimum Guaranteed Bit Rate bearers (GBR bearers):** GBR bearers provide communication services that need permanently allocated radio and network resources, such as VoIP. If additional resources are available to the users, the network may allow higher bit rates than the guaranteed bit rate, up to a Maximum Bit Rate threshold (MBR).
- **Non-GBR bearers:** non-GBR bearers do not support permanent radio and network resource allocation and, therefore, do not provide bit rate guarantees. Non-GBR bearers are normally used to provide best effort data services, such as web browsing, FTP transfer, instant messaging, etc.

2.4.1.2 QoS Class Identifiers

In order to guarantee uniform traffic handling across functional entities from different handset manufacturers and network vendors, 3GPP Technical Specifications define a standardised set of QoS Class Identifiers (QCIs) and Allocation and Retention Priorities (ARPs). The former determines how user-plane data packets will be handled in terms of priority, packet delay budget, and acceptable packet loss rate. The latter determines if a requested bearer should be established in case of radio congestion. Table 1 shows the standardised QCI values in the 3GPP EPS, as defined in [25].

The Packet Delay Budget metric (PDB) defines an upper bound for the time a user-plane data packet can be delayed between the UE and the PCEF. Specifically, the PDB represents the maximum packet delay with a confidence level of 98% in order to support the configuration of the scheduling and link layer functions. For a given QCI value, the corresponding PDB is the same for the uplink and downlink. Analogously, the Packet Error Loss Rate metric (PELR) defines an upper bound for the rate of upper-layer packets that have been processed by the sender of a link layer protocol (RLC) but have not been successfully delivered by the corresponding receiver to the upper layer (PDCP). Thus, the PELR defines an upper bound for a rate of non-congestion related packet losses. As in the case of packet delay budget, the PELR value is the same in the uplink and downlink. In terms of the QoS applicability domain (i.e., the EPS region where the QoS policies are enforced), the scope of the QCI characteristics for a client-server communication (e.g., UE to Internet server) comprises all EPS user-plane entities between the UE and the PCEF. In the case of peer-to-peer communications (UE to UE), the scope is extended from the PCEF to the destination UE. This is shown in Figure 16.

Table 1.- Standardised QCI values in the EPS

QCI	Type	Priority	Packet Delay Budget (ms)	Packet Error Loss Rate	Service Example
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	3	50	10^{-3}	Real-time gaming
4	GBR	5	300	10^{-6}	Non-conversational video (buffered streaming)
5	Non-GBR	1	100	10^{-6}	IMS signalling
6	Non-GBR	6	300	10^{-6}	Video (buffered streaming)
7	Non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
8	Non-GBR	8	300	10^{-6}	TCP-based (WWW, e-mail), chat, FTP, P2P file sharing, progressive video, etc.
9	Non-GBR	9	300	10^{-6}	Same as (8)

2.4.1.3 Traffic Flow Templates

The 3GPP Evolved Packet System uses IP packet filters known as *Traffic Flow Templates* (TFTs) in both the UE and P-GW functional entities to map user-plane data flows onto the corresponding EPS bearers, as shown in Figure 17. Thus, in the downlink, the P-GW injects packets into the corresponding EPS bearer according to the filtering rules defined in the downlink TFTs (DL TFTs). Analogously, the UE injects uplink data into the corresponding EPS bearer according to the rules defined in the uplink TFTs (UL TFTs). In general, uplink and downlink traffic flow templates filter data packets based on information contained in the IP header (such as source/destination IP addresses), as well as transport protocols and port numbers.

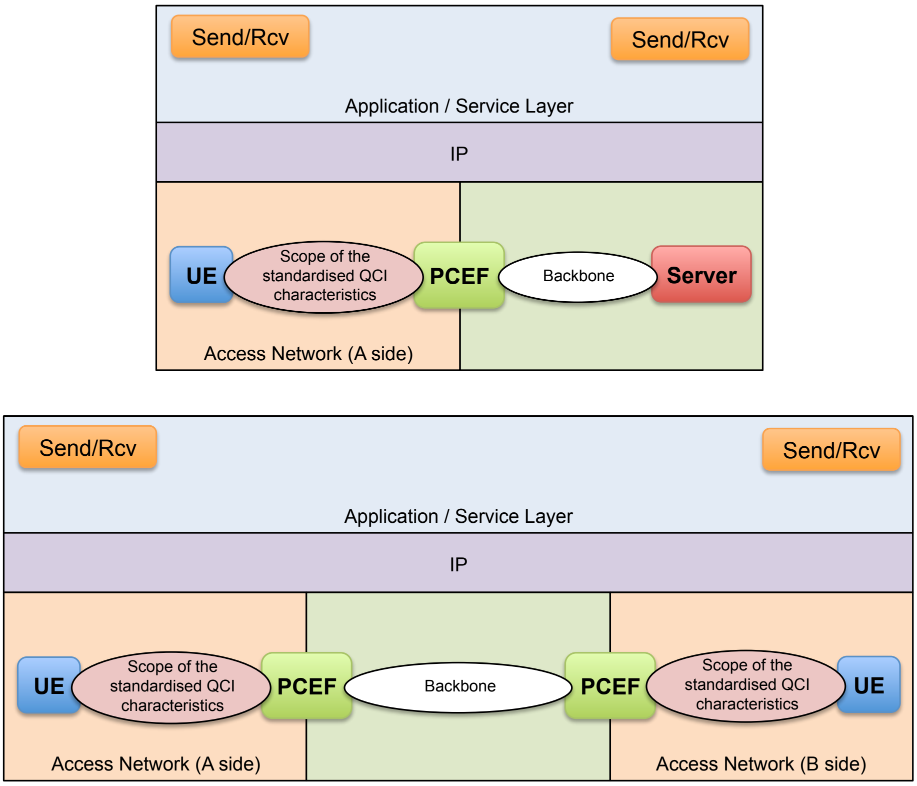


Figure 16.- Scope of the standardised QCI characteristics

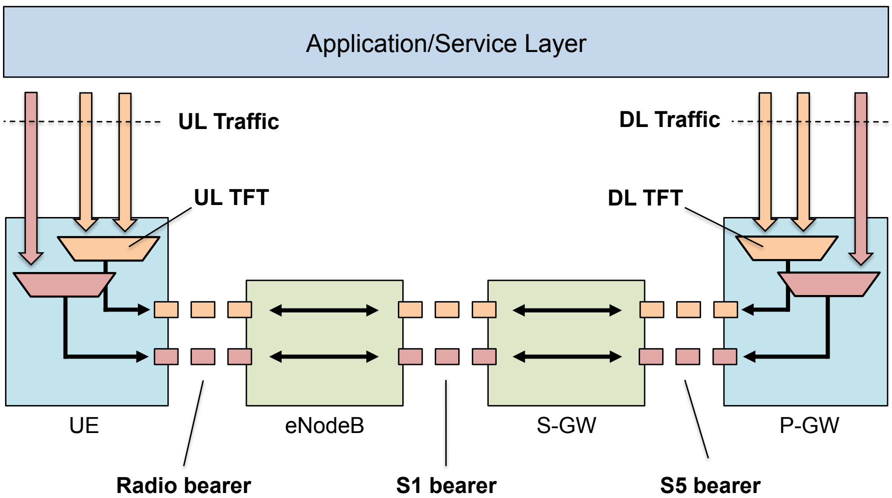


Figure 17.- Mapping between TFTs and {radio, S1, S5} bearers

2.4.1.4 The EPS Bearer Establishment Procedure

The EPS bearer is the key QoS-enforcement building block in the Evolved Packet System. When a UE (or the EPC) initiates the establishment of an EPS bearer, functional entities in the Evolved Packet System trigger the establishment of the subsequent S5, S1, and radio bearers. In this subsection we describe the 3GPP network procedure aimed at establishing an EPS bearer between a UE and the P-GW.

When a UE attaches to an LTE network for the first time, the EPS assigns an IP address to the UE and establishes a *default EPS bearer* between the UE and the P-GW. This bearer is referred to as the *default bearer* in the sense that it provides the UE with always-on IP connectivity with the P-GW and, in turn, an external network. Default bearers are always non-GBR bearers, as they remain active throughout the lifetime of a connection with the external network. On the other hand, the UE and the EPC can respectively request/initiate the establishment of one or more *dedicated EPS bearers*. Dedicated bearers sit on top of the default EPS bearer in order to support differentiated QoS treatment for specific user-plane traffic flows between the UE and the P-GW, such as voice over LTE (VoLTE), real-time gaming, videoconferencing, etc. In terms of bit rate, dedicated bearers can be either GBR or non-GBR, depending on the service provided. In both cases, the EPC must execute the necessary network procedures to ensure that the corresponding S5, S1, and radio bearers are being properly established. Figure 18 shows the message sequence chart of the 3GPP UE-initiated EPS bearer establishment procedure. The steps in the figure are described below:

- **Steps 1 – 2:** the UE initiates a Random Access procedure to allocate a set of dedicated radio resources with the eNB. A detailed description of the LTE Random Access procedure can be found in [24].
- **Step 3:** the UE sends an *RRC Connection Setup Complete* message to the eNB to confirm the establishment of the RRC connection. In addition, this message encapsulates a *NAS Bearer Resource Allocation* message to instruct the MME that the UE is requesting the establishment of an EPS bearer.
- **Step 4:** the eNB forwards the *NAS Bearer Resource Allocation* message to the MME encapsulated in an *S1-AP Initial UE* message.
- **Step 5:** the MME processes the *NAS Bearer Resource Allocation* message and sends a *GTP-C Bearer Resource Command* message to the S-GW. The purpose of this message is to instruct the P-GW to initiate the establishment of an EPS bearer with the UE.
- **Step 6:** the S-GW forwards the *GTP-C Bearer Resource Command* to the P-GW.

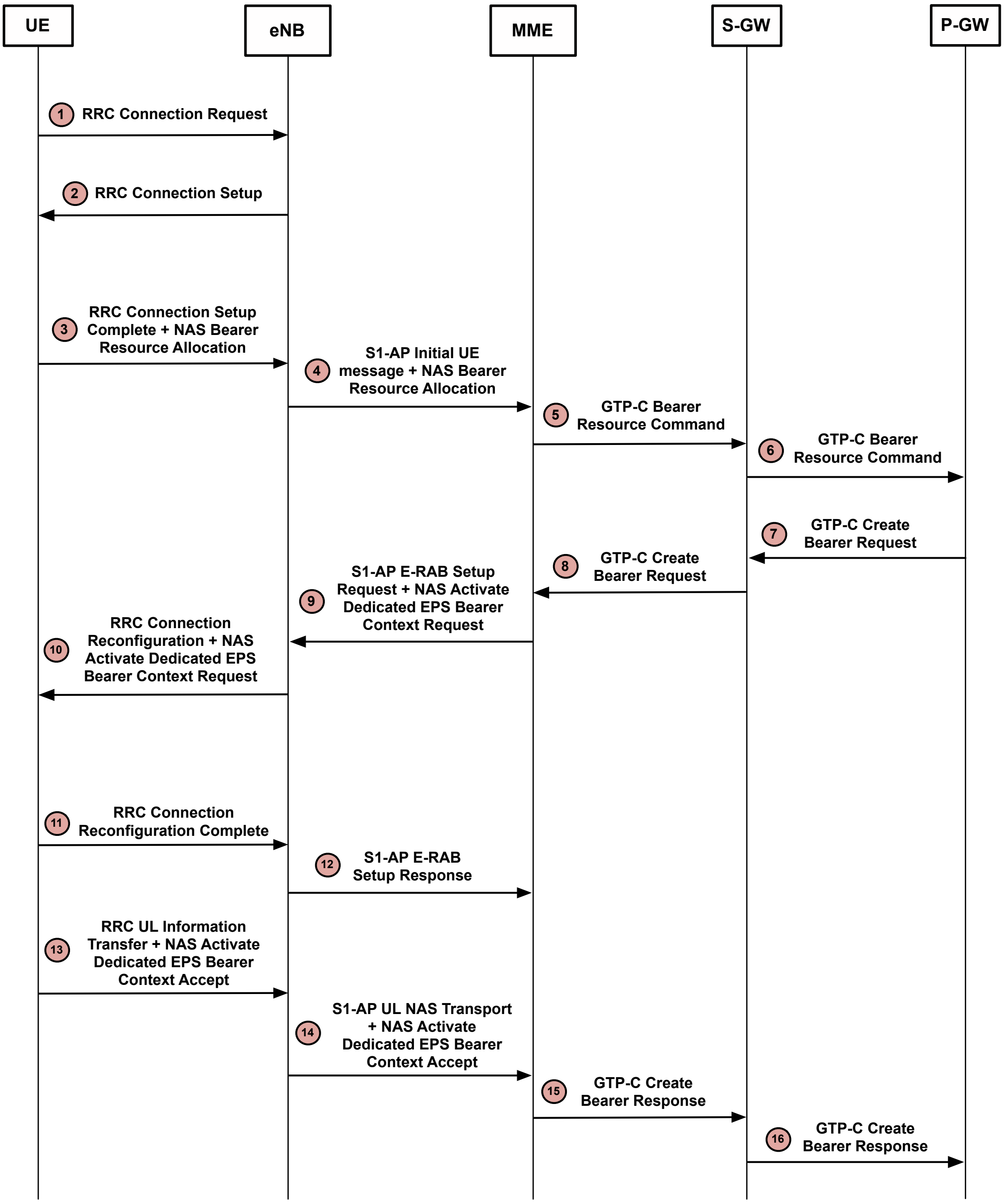


Figure 18.- The 3GPP UE-initiated EPS bearer establishment procedure

- **Step 7:** the P-GW sends a *GTP-C Create Bearer Request* message to the S-GW. The purpose of this message is to initiate the establishment of an S5 bearer between the P-GW and the S-GW. Note that this message, in turn, will trigger the establishment of the S1 and radio bearers (i.e., the E-RAB).

- **Step 8:** the S-GW transparently forwards the *GTP-C Create Bearer Request* message to the MME.

- **Step 9:** the MME sends an *S1-AP E-RAB Setup Request* message to the eNB to initiate the establishment of an S1 bearer. In addition, this message contains a *NAS Activate Dedicated EPS Bearer Context Request* message in order to configure a radio bearer between the UE and the eNB.

- **Step 10:** the eNB sends an *RRC Connection Reconfiguration* message to the UE. This message contains the necessary configuration parameters (i.e., PDCP, RLC, MAC, and L1 settings) to initialise the LTE-Uu protocol stack. In addition, it also contains a *NAS Activate Dedicated EPS Bearer Context Request* message. The purpose of this message is to establish an EPS bearer context with a specific QoS and TFT at the UE.

- **Step 11:** the UE acknowledges the initialisation of the LTE-Uu interface with an *RRC Connection Reconfiguration Complete* message.

- **Step 12:** the eNB sends an *S1-AP E-RAB Setup Response* message to the MME to confirm the successful establishment of a radio bearer between the UE and the eNB.

- **Step 13:** the UE sends a *NAS Activate Dedicated EPS Bearer Context Accept* message encapsulated in an *RRC UL Information Transfer* to the eNB. The purpose of this message is to confirm the successful establishment of an EPS bearer context in the UE.

- **Step 14:** the eNB forwards the *NAS Activate Dedicated EPS Bearer Context Accept* message to the MME encapsulated in an *S1-AP UL NAS Transport* message.

- **Step 15:** the MME sends a *GTP-C Create Bearer Response* message to the S-GW in order to confirm the successful establishment of an E-RAB between the UE and the S-GW.

- **Step 16:** the S-GW sends *GTP-C Create Bearer Response* to confirm the successful establishment of the S5 bearer (and, in turn, the EPS bearer) to the P-GW.

2.4.1.5 Local IP Access and Selected IP Traffic Offload

In the EPS bearer service architecture, traffic exchange between two peer entities (e.g., two UEs, a UE and an Internet server, etc.) must necessarily involve the establishment of an EPS bearer between the UE and the P-GW. Thus, when a UE wants to establish a connection with a peer entity in an external network, it must first initiate an EPS bearer with the P-GW, as the P-GW is the EPS functional entity that acts as a gateway node between the EPC and external networks. However, this approach is not particularly efficient in small cell scenarios where the destination node might be a local host connected to the same local IP network as the source node. In these cases, the UE would first have to establish an EPS bearer with the P-GW in the EPC. Then, the P-GW would have to establish a subsequent EPS bearer with the destination node in the local IP network in order to support traffic exchange between the source and destination nodes. Thus, the source UE will end up routing data packets to the destination UE via the P-GW in the EPC. This cumbersome routing scheme is referred to as *trombone routing*.

In a similar way, a UE might want to communicate with a peer entity located in an external network that is directly accessible through the backhaul connection of the serving small cell (e.g., an Internet server), thus effectively bypassing the mobile operator's network. In this scenario, the UE also would also need to traverse the MNO core network in order to reach the P-GW and, in turn, the host in the external network. In order to avoid a waste of RAN and core network resources, Release 9 and 10 of 3GPP Technical Specifications have defined two new modes of operation for small cells: Local IP Access (LIPA) and Selected IP Traffic Offload (SIPTO), respectively.

LIPA provides access for IP-capable UEs camped on a small cell to other IP-capable entities (e.g., a local server) in the same residential/enterprise IP network. Thus, LIPA eases congestion in the small cell backhaul and the EPC by steering local traffic via the serving HeNB. Analogously, SIPTO enables selected IP traffic offloading towards external networks (e.g., Internet) via the HeNB. Thus, selected IP packets reaching the HeNB will be offloaded towards the Internet through the small cell backhaul in order to alleviate congestion in the mobile operator's Evolved Packet Core. A basic example of LIPA and SIPTO traffic flows is shown in Figure 19.

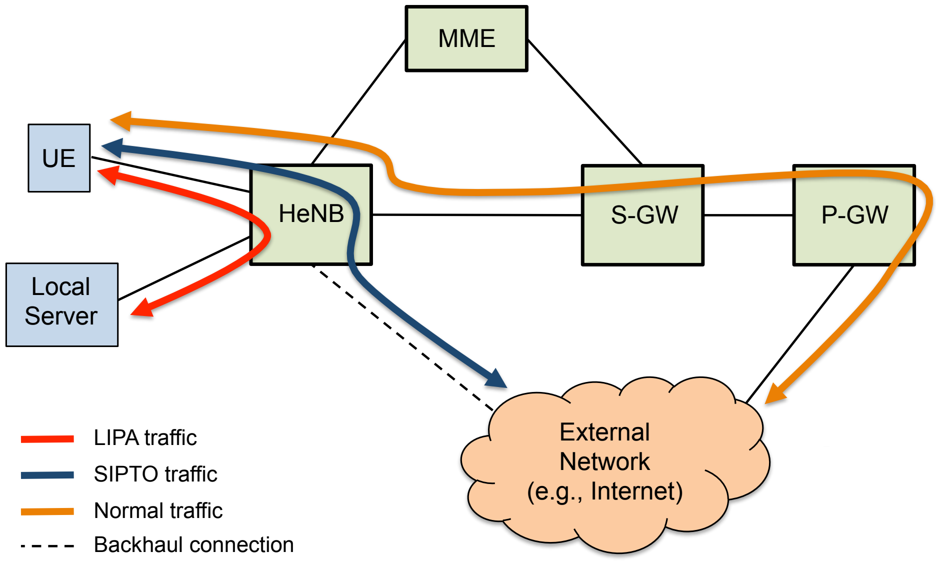


Figure 19.- LIPA and SIPTO communication scenarios

3GPP Technical Specifications define three different approaches for LIPA/SIPTO, namely (a) a solution based on a dedicated offload PDN connection (whereby separate PDN connections are used for LIPA/SIPTO and non-LIPA/SIPTO traffic), (b) a solution based on Network Address Translation (NAT) and a Traffic Offload Function (TOF), where a single PDN connection is used for both LIPA/SIPTO and non-LIPA/SIPTO traffic, and the breakout functionality including NAT is collocated in the HeNB; and (c) a Local Gateway-based (L-GW) architecture, which is practically identical to (a) with the exception of an additional extension tunnel being established between the HeNB in the local IP network and the P-GW in the EPC (note that the L-GW is a functional entity that implements some of the functions of the P-GW in the local IP network). These variants are described in detail in [23].

In terms of standardisation, 3GPP selected the first approach as the architectural framework for supporting LIPA/SIPTO services in the Evolved Packet System. As described in [23], this solution requires either a standalone (or collocated) L-GW deployment in the local network or HeNB, respectively. Furthermore, it introduces minimal standardisation changes compared to the other two approaches, as it does not involve interface modifications, packet inspection, or NAT.

2.4.2 Multicast Traffic Management

In communication networks, the term *multicast* refers to the distribution of data to an audience of multiple users [24]. Instead of sending identical packets to each node in a unicast manner, a multicast/broadcast node transmits a single user-plane data flow to all participating receivers (broadcast) or subset of receivers (multicast), thus allowing a more efficient utilisation of the network resources. This is particularly convenient in wireless and cellular networks due to the scarcity of radio resources.

The *Multimedia Broadcast/Multicast Service* (MBMS[3], in short) is the standard 3GPP mechanism to provide broadcast and multicast services in the Evolved Packet System. MBMS for LTE was first introduced in Release 9 of 3GPP Technical Specifications, although previous specifications for UMTS systems were also defined in Release 6. MBMS supports multiple multicast/broadcasts services, such as live video and audio streaming, push media, e-publications, Operating System (OS) updates, newsreels, weather and traffic forecasts, etc.

MBMS also defines two modes of operation, namely the *single-cell* and *multi-cell* transmission modes. In single-cell mode, all eNBs participating in the MBMS service use the same frequency and time resources to transmit the same MBMS data in a tightly synchronised manner. As a result of this, all eNBs in single-cell mode data appear to the receiving UE as a single MBMS cell. This is also known as *Multicast/Broadcast Single Frequency Network* (MBSFN). In multi-cell mode, each eNB transmits one or more MBMS sessions independently, i.e., in an uncoordinated fashion with respect to other eNBs.

In the following subsections we provide an overview of the MBMS network architecture in the 3GPP Evolved Packet System, along with its corresponding user-plane and control-plane protocol stacks. For further information on the MBMS service the reader is referred to [7].

2.4.2.1 The MBMS Network Architecture

Figure 20 shows the 3GPP functional entities and reference points in the MBMS network architecture. These functional entities are described below:

---

[3] The term *Evolved Multicast/Broadcast Multimedia Service* (eMBMS) is also used in the literature to differentiate between Release 6 MBMS (UMTS) and Release 9 MBMS (LTE).

- **Broadcast/Multicast Service Centre (BM-SC):** the BM-SC is in charge of scheduling an MBMS service (e.g., a live video stream), announcing the service to UEs in the E-UTRAN, authorising users, allocating bearer service identification, and initiating/terminating MBMS bearer resources. Optionally, it may also be the direct interface with the EPS for over-the-top content providers.

- **MBMS Gateway (MBMS GW):** this entity is the entry point for MBMS control- and user-plane traffic in the EPC. The MBMS GW uses IP multicast to distribute MBMS user-plane traffic to all eNodeBs in the E-UTRAN. In addition, the MBMS GW also performs MBMS session control signalling (session start and stop) towards the E-UTRAN via the MME.

- **Multi-Cell/Multicast Coordination Entity (MCE):** this functional entity is in charge of managing the necessary radio resources for multicast/broadcast traffic distribution. Some of its functions comprise allocating radio resources and performing admission control for all eNBs participating in the MBMS service, selecting the optimal modulation and coding scheme (MCS) to maximise the quality of experience (QoE) for all users, and establishing and releasing MBMS sessions in the E-UTRAN.

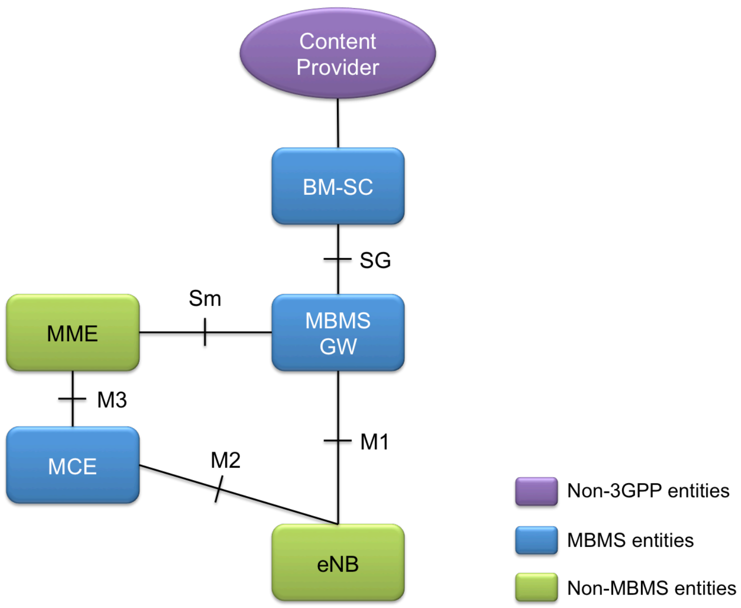


Figure 20.- The MBMS network architecture

3GPP Technical Specifications allow the deployment of MCEs as standalone network elements in the Evolved Packet System architecture or, alternatively, collocated with an eNodeB. This is shown in Figure 21. In the leftmost figure, a single MCE has been deployed as a standalone element in the E-UTRAN architecture. In the rightmost figure, individual MCEs have been collocated in each eNB. As far as the MBMS GW is concerned, this functional entity is split in two sub-entities, namely the MBMS CP (control-plane) and the MBMS UP (user-plane). These sub-entities communicate with the BM-SC via the SG-mb and SG-imb interfaces, respectively.

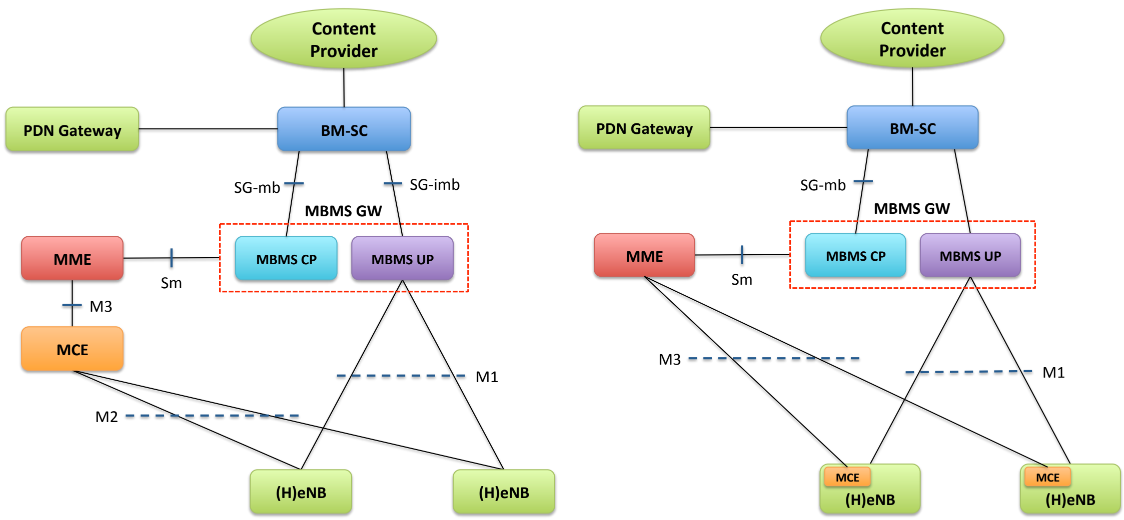


Figure 21.- Standalone and collocated MCEs in MBMS

The 3GPP interfaces in the MBMS network architecture are described below:

- **M3 (MME – MCE):** this interface supports the exchange of the M3 Application Part control-plane protocol (M3-AP) between the MME and the MCE. The M3-AP protocol defines network procedures for starting, stopping, and updating MBMS sessions. When an MBMS session needs to be started/updated, the MME provides details about the corresponding MBMS bearer whilst the MCE verifies whether the multicast/broadcast service can be supported. The M3 is a pure control-plane interface.

- **M2 (MCE – eNB):** the M2 interface supports the exchange of M2 Application Part control-plane PDUs (M2-AP) between the MCE and the eNB for starting, stopping, and updating MBMS sessions. In particular, the MCE functional entity provides the corresponding radio resource configuration to all eNBs participating in an MBMS session. Later on, eNBs broadcast this configuration to all UEs in the session. Note that, just like the M3 interface, the M2 interface is a pure control-plane reference point.

- **M1 (MBMS GW – eNB):** the M1 interface supports the exchange of MBMS data streams between the MBMS GW and all eNBs participating in the MBMS session. IP multicast is used for point-to-multipoint delivery of user-plane data packets. As opposed to M2 and M3, the M1 interface is a pure user-plane interface.
- **Sm (MME – MBMS GW):** the Sm reference point is a pure control-plane interface for the exchange of service control messages between the MME and the MBMS GW (e.g., the IP multicast address for MBMS data reception).
- **SG-mb (BM-SC – MBMS GW):** the SG-mb interface provides MBMS bearer service-specific signalling, such as session start, update, and stop between the BM-SC and the MBMS-GW. Amongst other, this includes session attributes like the QoS profile and MBMS service area (i.e., the region of the cellular network where the MBMS service will be broadcasted).
- **SG-imb (BM-SC – MBMS GW):** this is a pure user-plane reference point between the BM-SC and MBMS GW for MBMS data delivery.

2.4.2.2 The MBMS Protocol Stack Architecture

Figure 22 shows the M1 interface user-plane protocol stack in the MBMS network architecture. The communication protocols that do not specifically belong to the M1 interface are shown in grey. As seen in the figure, the BM-SC sends the MBMS data packets to the UEs participating in the MBMS session via the MBMS GW and, in turn, the eNB. Additionally, the SYNC protocol between the BM-SC and the eNB adds timestamps to MBMS data in order to support content synchronisation, radio frame transmission identification, and packet loss detection. For a detailed description of the SYNC protocol the reader is referred to [26]. Note that the transport network layer of the M1 interface is the same as that of the S1-U and X2-U interfaces (i.e., GTP-U over UDP over IP).

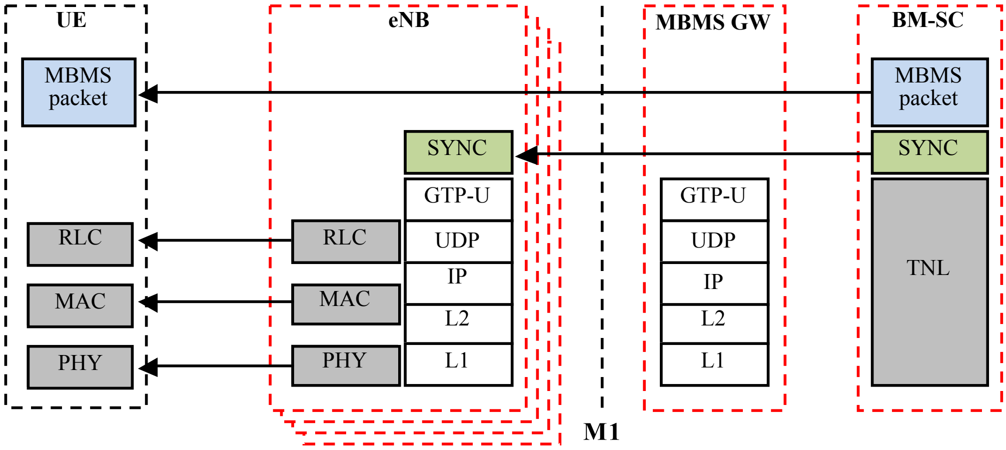


Figure 22.- The M1 user-plane protocol stack

Figure 23 shows the control-plane protocol stack for M2 and M3 interfaces in the MBMS network architecture. The communication protocols that do not specifically belong to the M2 and M3 interfaces have been greyed out. As previously discussed, the M2-AP and M3-AP protocols are responsible for the exchange of MBMS control-plane information (session handling, scheduling provision, session suspension and resumption, etc.) between the eNB and MCE, and MCE and MME, respectively. Note that, as in the S1-MME and X2-C interfaces, the transport network is based on SCTP over IP. For a detailed description of the M2-AP and M3-AP protocols the reader is referred to [27] and [28].

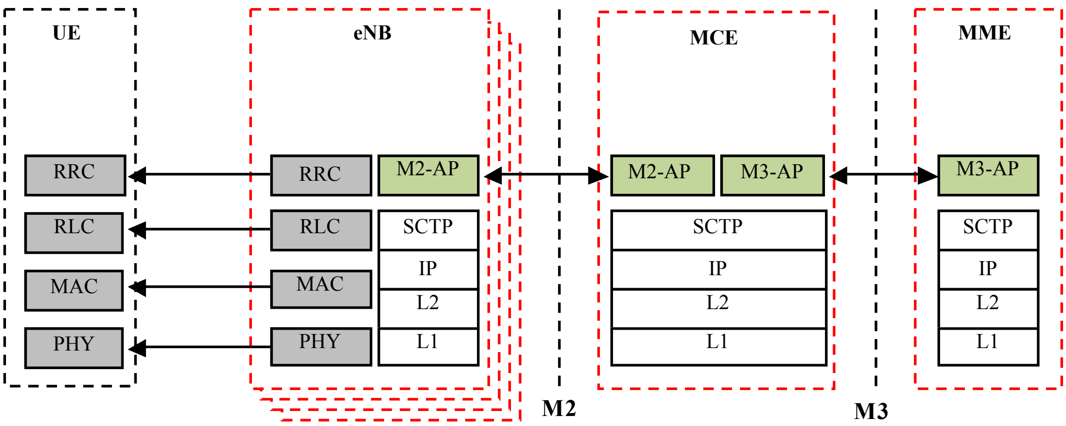


Figure 23.- The {M2, M3} control-plane protocol stacks

2.4.2.3 The MBMS Session Start Procedure

Figure 24 shows the message sequence chart of the 3GPP MBMS Session Start network procedure. The aim of this procedure is to allocate the necessary E-UTRAN and EPC resources to start an MBMS session. Note that, in LTE networks, there is no specific support for multicast group subscription, i.e., MBMS sessions are announced to all UEs via broadcast channels and, later on, individual users decide if they want to join the session. In general, MBMS service announcement and joining/leaving procedures are application-specific and, thus, beyond the scope of 3GPP network procedures. The steps in Figure 24 are described below:

- **Step 1:** the BM-SC sends a *Session Start Request* message to the MBMS GW in order to initiate an MBMS session. Then, the MBMS GW forwards this message to the MME over the Sm interface, along with the IP multicast transport layer address and the downlink tunnel endpoint ID (TEID) for the M1 transport association.
- **Step 2:** the MME sends an *M3-AP MBMS Session Start Request* message to the MCE with the configuration parameters for the MBMS session. These settings contain the source and destination IP multicast addresses, QoS requirements, MBMS service area identification, session duration, etc.
- **Steps 3 – 4:** upon reception of the *M3-AP MBMS Session Start Request* message, the MCE performs admission control functions and configures the radio resources in the MBMS eNBs. The MBMS session and radio resource parameters are sent to eNBs in the *M2-AP MBMS Session Start Request* and *M2-AP MBMS Scheduling Information* messages, respectively.
- **Steps 5 – 6:** each eNB participating in the MBMS session sends an *M2-AP MBMS Session Start Response* message to the MCE to confirm the reception of the MBMS session configuration. Likewise, each eNBs sends an *M2-AP MBMS Scheduling Information Response* message to the MCE to acknowledge the reception of the MBMS radio resources.
- **Steps 7 – 8:** eNBs notify a change in the MBMS control information over the Multicast Control Channel (MCCH). Immediately after, each eNB sends the new MBMS radio resource configuration information in the *RRC MBSFN Area Configuration* message to all participating UEs. Finally, MBMS eNBs join the transport network multicast IP address and the transfer of MBMS user-plane data is initiated.

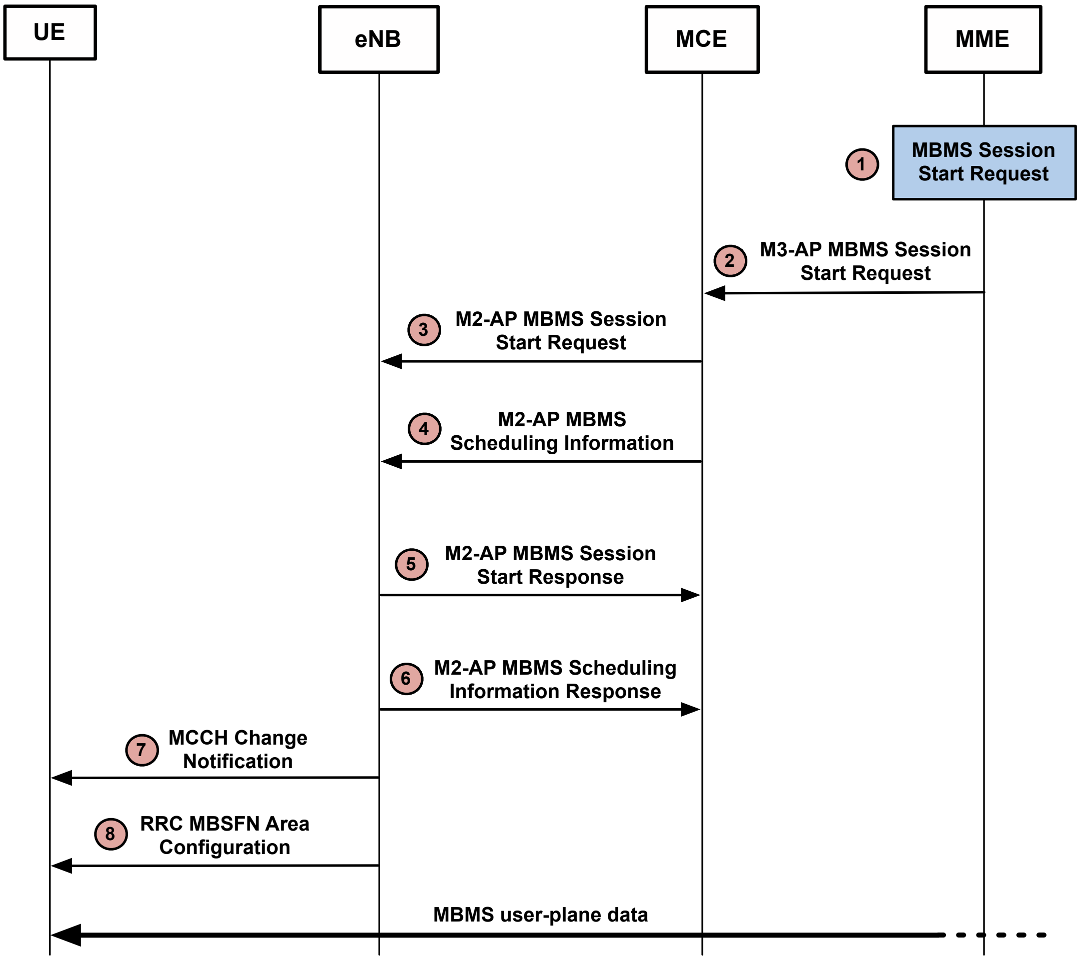


Figure 24.- The 3GPP MBMS Session Start procedure

Chapter 3

State of the Art

In this chapter we discuss some of the most recent developments in the three main research areas of this Ph.D. Thesis, namely location management, traffic management, and new architectures for large-scale networks of small cells. The aim of this chapter is to provide the reader with a general overview of the state of the art in each of these topics from the research and 3GPP standardisation point of view. Thus, each subsection includes up-to-date references to contributions in the literature, as well as to the corresponding 3GPP Technical Specifications, whenever applicable.

3.1 New Architectures for Networks of Small Cells

Since the introduction of small cells in Release 8 of 3GPP Technical Specifications, research work on this topic has primarily focused on standalone deployments for residential and small office scenarios, such as the ones described in Section 2.2.1. Some of the most common research topics in small cells have traditionally been radio resource allocation, interference management, access control, self-organising capabilities, timing and synchronisation, QoS provision over the small cell backhaul, handover optimisation, power control, security, etc. [2]. As far as the network architecture is concerned, the latest release of the 3GPP Technical Specifications (at the time of writing, Release 12) only supports standalone small cell deployments [7]. Thus, the design of a scalable network architecture capable of supporting massive deployments of wired/wireless networks of small cells in the 3GPP Evolved Packet System still remains an open issue. In the following subsections we provide an overview of some related work on new architectural frameworks for enhanced LTE networks at both the standardisation (i.e., 3GPP) and research levels.

3.1.1 3GPP Technical Specifications

The concept of networks of small cells was originally introduced in the context of the European Project BeFEMTO [4]. Some of the research results obtained during the course of this project are currently undergoing standardisation discussions in 3GPP as part of the *LIPA Mobility and SIPTO at the Local Network* work item (LIMONET) [29]. The aim of this standardisation effort is to define a 3GPP-compliant architectural and procedural framework to support local traffic in standalone small cell deployments in the Evolved Packet System. Although the concept of networks of small cells is gaining momentum in both the industry and the academia, to the best of our knowledge there is no previous standardisation work on architectural solutions capable of supporting large-scale deployments of NoS in the Evolved Packet System.

3.1.2 Industrial and Academic Research

Most of the industrial and academic research activity in new architectural frameworks for cellular networks is focused on emerging technologies such as *Cloud RAN*[4] (C-RAN), *RAN as a Service* (RANaaS), and *Software-Defined Evolved Packet Core* (SD-EPC). Though these technologies do not strictly address the issue of defining a 3GPP-compliant architecture to support massive deployments of networks of small cells in the Evolved Packet System, they constitute a promising step towards the realisation of enhanced and more efficient cellular networks. In addition, the standardisation effort carried out by 3GPP to define the Evolved Packet System in terms of pools of functional entities (therefore decoupling the actual functionality from the underlying network elements) facilitates its portability to software-defined network scenarios. In the following subsections we provide a brief overview of the state of the art in each one of these areas (C-RAN, RANaaS, and SD-EPC).

3.1.2.1 Cloud-RAN Based Cellular Networks

In recent years, cloud-based radio access networks have gained attention from mobile network operators and network vendors [30], [31], [32] as a mechanism to harness the capacity benefits of smaller base stations and reduce capital and operational expenditures. Essentially, a C-RAN architecture is based on three main components, namely (a) a set of distributed Radio Access Units (RAUs), each deployed at the corresponding remote cell; (b) a pool of Baseband Units (BBUs) located in a datacentre and running on high-performance processors through real-time virtualisation techniques, and (c) a high-bandwidth, low-latency optical transport

---

[4] In the literature, the acronym C-RAN refers to the terms *Cloud RAN* and *Centralised RAN* indistinctively.

network to provide connectivity between BBUs and RAUs [33]. In this scenario, RAUs act as lightweight radio units, with the processing block being fully migrated to the BBU. By decoupling RAUs from BBUs, network vendors can achieve lower costs for RAN equipment, thereby enabling large-scale deployments of smaller cellular base stations in localised and high user-density areas. On the other hand, centralised processing in BBUs leverages the advantages of high-performance datacentres (such as dynamic allocation of computing resources, high processing power, virtualisation, etc.) in order to achieve efficient centralised radio resource management.

The authors in [33] provide a comprehensive analysis of the state of the art in C-RAN technologies in the industry. According to the authors' claims, in recent years network vendors have primarily focused on developing more scalable and power-efficient RAN equipment (e.g., Alcatel-Lucent's LightRadio [31], Nokia Solutions and Networks' Liquid Radio [32], or Ericsson's AIR [34], amongst others) in order to support multiple frequency bands and radio access technologies (UMTS, LTE, LTE-A, etc.). On the other hand, chipset manufacturers like Intel, Broadcom or Texas Instruments have mainly worked on improving the low-level baseband architecture. Whilst some of these companies support the use of general-purpose processors for implementing intensive baseband functions [35], other advocate for application-specific DSPs [36]. More recently, computer architecture designer ARM (traditionally associated with power-efficient virtual architectures for mobile and portable devices) has started to develop products for the networking and server markets [37]. Finally, the issue of optical backhaul is also discussed in the paper. Optical backhaul links between RAUs and BBUs are already being used in several Distributed Antenna Systems (DAS) for stadiums, convention centres, airports, etc. In these systems, radio signals can be transported either in analogue (e.g., radio-over-fibre) or digital (e.g., Common Public Radio Interface (CPRI)) modes. In fact, the provision of ubiquitous low-latency, high-throughput optical links between RAUs and BBUs remains one of the bigger obstacles for the large-scale rollout of cloud-based radio access networks.

3.1.2.2 Radio Access Network as a Service

The authors in [38] discuss the concept of Radio Access Network as a Service, i.e., an extension of the C-RAN paradigm where full RAN blocks can be dynamically allocated (and moved) from RAUs to a cloud server infrastructure, and vice-versa.

Essentially, in the RANaaS paradigm, standalone protocol layers in the LTE-Uu control- and user-plane protocol stacks (e.g., MAC, RLC, PDCP, RRC, etc.) are dynamically allocated and initialised either on the radio access units or the cloud servers according to specific traffic demands and available backhaul resources. This approach eases the latency and throughput requirements on backhaul links between RAUs and the cloud infrastructure. This is achieved by running critical baseband tasks in radio access units, whilst moving less time-sensitive operations to the datacentre infrastructure. Thus, whilst the C-RAN paradigm centralises all baseband functions in cloud-based BBUs, RANaaS provides a flexible framework where radio resources and protocol entities move dynamically between RAUs and the datacentre infrastructure. In addition, RANaaS leverages generic Infrastructure as a Service models (IaaS), thus enabling the use of virtualisation, commercial off-the-shelf operating systems, and middleware enterprise software to achieve a more efficient usage of datacentre resources and lower network infrastructure costs. This is particularly interesting for mobile network operators, as they can benefit from reduced CAPEX/OPEX and more flexible cellular network deployments.

The authors in [39] propose an RANaaS-based architecture for cellular networks based on four key enablers: (a) a large-scale deployment of small cellular base stations as a way to improve network coverage and boost capacity for end users, (b) the use of centralised radio resource management algorithms to implement more efficient inter-cell interference avoidance/mitigation schemes and optimise radio and network resources usage, (c) the use of RANaaS techniques as a flexible trade-off between fully-centralised (C-RAN) and traditional EPS deployments, and (d) a joint design of the RAN and backhaul in order to achieve a more efficient cellular network operation. In summary, the paper points at the issue of backhaul resource availability as one of the key challenges in future cellular networks. The authors' proposal exploits RANaaS schemes to provide dynamic radio and network resource allocation between RAUs and datacentre servers based on the availability of backhaul resources. Thus, if enough backhaul resources are available, more functionality is deployed in a centralised way. Instead, if the cellular network is backhaul-limited (e.g., wireless backhaul), more functional blocks are pushed from the datacentre infrastructure to the RAUs.

3.1.2.3 Software-Defined Evolved Packet Core

A Software-Defined Evolved Packet Core is the realisation of the 3GPP EPC through network function virtualisation techniques (i.e., the implementation of network functions in standalone software modules that can run on a range of industry-standard server hardware). In SD-EPCs, functional entities such as the MME, the HSS, the PCRF, the S-GW, or the P-GW are fully implemented in software running on high-performance datacentre servers. Mobile network operators can benefit from the SD-EPC paradigm, as deploying network functions as virtualised instances instead of application-specific hardware reduces CAPEX/OPEX and enables more efficient network configuration, operation, and maintenance schemes.

The authors in [40] propose an architectural solution for the deployment of a Software-Defined 3GPP Evolved Packet Core. In particular, they define a functional split in the EPC in order to create a centralised datacentre and an operator's transport network domain. In addition, they define four alternative deployment frameworks based on software-defined networks (SDNs) according to the mapping of network functions to the datacentre and/or transport network domains. These alternatives comprise (a) a Full Cloud Migration Architecture (i.e., the full virtualisation of the MME, S-GW, and P-GW network functions), (b) a Control-plane Cloud Migration Architecture (only the MME, S-GW, and P-GW control-plane network functions are virtualised), (c) a Signalling Control Cloud Migration Architecture (only the signalling function for all EPC entities is virtualised), and (d) a Scenario-based Migration Architecture (all network functions are implemented in both domains simultaneously, although only one implementation is enabled at each time depending on specific traffic demands). Although the authors provide a detailed discussion of each one of these options, finding the optimal deployment solution via a quantitative evaluation of the deployment split is left for future work.

The authors in [41] provide a comprehensive analysis of the roles and functions of traditional and virtual mobile network operators (MVNOs) and how they can leverage cloud-based technologies to address the challenges of future cellular networks. In particular, the paper discusses some of the limitations of the SD-EPC paradigm, namely (a) the illusion of infinite resources (i.e., how cloud models that are viable for utility computing might not be directly applicable to cellular networks due to external factors like the lack of proper radio coverage and backhaul availability in rural areas), (b) the elimination of upfront commitment by cloud users (the question of who would provide the network and computing infrastructure for SDN-based cellular networks (MNOs, third-party cloud computing providers (e.g., Amazon, Google), public-private partnerships, etc.)), and (c) the ability to pay for allocated resources on a short-term basis (i.e., the implementation of efficient RAN and EPC pay-as-you-use schemes instead of allocating an arbitrary amount of radio and network resources to SD-EPC users). Furthermore, the authors outline some of the technical issues that must be addressed to realise the so-called utility cellular networking concept. These comprise the efficient design of SDN-based shared cellular networks, the implementation of joint RAN/EPC resource allocation algorithms, and the impact of a highly virtualised network infrastructure on future 3GPP Technical Specifications.

3.1.3 Conclusions

Networks of small cells are gaining attention at both the academic and industrial level as a solution to the issue of capacity crunch in current (and future) cellular networks. Alternative approaches comprise Cloud RAN, RAN as a Service, and a Software-Defined Evolved Packet Core. Though promising and technically sound, none of these proposals provides a 3GPP-compliant architecture to support massive deployments of small cells (or, in a wider sense, small cellular base stations) in the context of the Evolved Packet System.

3.2 Location Management in Cellular Networks

As described in Section 2.3, location management is a collective term in cellular networks that comprises both the paging and Tracking Area Update procedures. Traditionally, standard 3GPP location management mechanisms have been defined in the context of macrocell networks. This has implications in their performance in large-scale networks of small cells. Thus, the research work carried out in this Ph.D. Thesis focuses on location management procedures *within* the scope of the NoS domain. This is commonly known as *Local Location Management* (LLM). LLM functions are in charge of keeping track of the location of UEs in the customer domain of a network of small cells. Their objective is to determine the serving HeNB where the destination UE is camped on whenever an incoming voice call or data connection needs to be established.

The following subsections discuss some of the recent research work on Tracking Area Update and paging procedures for cellular networks in general, i.e., without making a particular case for networks of small cells. Although most of the state of the art is topology- and RAT-agnostic (i.e., it can be applied to different network deployments (macrocells vs. NoS) and cellular technologies (GSM, UMTS, LTE, etc.)), we have specifically considered research contributions that discuss location management mechanisms in the context of the Evolved Packet System.

3.2.1 Tracking Area Update

A large amount of previous research work has been carried out in the area of adaptive location management in cellular networks [42]–[48]. However, most of these proposals address the issue of location management purely as an optimisation problem and, in general, do not consider the implications of implementing these mechanisms on commercial networks (e.g., peaks of spurious location-related signalling traffic throughout the network due to global tracking area reconfigurations). In addition, some of these schemes are not 3GPP-compliant, which further compromises their implementation in commercial networks.

In [46] and [47], the authors propose a mechanism to reduce location signalling traffic in LTE networks using the Tracking Area List (TAL) feature. Tracking Area Lists were introduced in 3GPP Release 8 to allow the Evolved Packet Core to register UEs to more than one tracking area simultaneously [5]. This way, the core network prevents UEs from performing bulk TAU procedures (e.g., the so-called *train scenario*, where a large number of UEs enter a new tracking area simultaneously). Both mechanisms deliver an optimised TAL configuration to all UEs in the network according to signalling computations made by individual cells in the radio access network. Whilst technically sound, this proposal has some important caveats. First, they cannot guarantee a global minimum for location signalling traffic, as each cell makes autonomous TAL configuration decisions without taking into account the impact of global TAL reconfigurations on neighbouring cells. Secondly, delivering periodical TAL reconfigurations to all UEs in the cellular network generates a substantial amount of spurious location signalling traffic. In practice, this means that a large number of UEs may need to perform a Tracking Area Update procedure every time the network delivers a new optimised TAL configuration that differs significantly from the previous one. Thus, the applicability of these adaptive location management mechanisms to commercial networks might be called into question.

The authors in [48] propose an adaptive location management mechanism to register UEs to dynamic TAs in order to reduce location signalling traffic in the cellular network. This mechanism derives the optimal cell configuration in the forthcoming tracking area from live UE speed and call arrival rate measurements. This way, UEs are required to send near real-time measurements to EPC functional entities every time they enter a new TA. Both the network and the UEs estimate the call arrival rate by averaging the number of incoming calls during a period of time. However, providing the core network entities with accurate UE speed measurements is not straightforward. To solve this problem, the authors propose two methods:

- UEs measure their speed using a Global Positioning System (GPS) receiver.
- UEs estimate their speed by combining internal time measurements with geographical coordinates advertised by the old and new base stations.

The first method is not realistic for various reasons. First, GPS signals might not always be sufficiently accurate (nor available), especially in indoor and dense urban scenarios. And, secondly, UEs move into RRC idle mode precisely to reduce power consumption and extend battery life, which is not consistent with keeping a GPS receiver locked and tracking satellite signals on a continuous basis. On the other hand, the second method is particularly difficult to implement in a network of small cells, as HeNBs might not always be able to advertise their geographic coordinates (e.g., indoor deployments, non GPS-enabled small cells, mobile small cells, etc.). Furthermore, the mechanism described in [48] is based on network-wide Tracking Area reconfigurations, such as those described in [46] and [47]. As previously discussed, these mechanisms involve a substantial amount of spurious location-related signalling traffic throughout the cellular network during global TA reconfigurations, which defeats the purpose of adaptive location management.

With regards to standardisation work in 3GPP, the standard Tracking Area Update procedure for LTE networks has been defined in [5]. To the best of our knowledge, no new 3GPP work items aimed at improving the TAU procedure are expected to be included in the latest release of 3GPP Technical Specifications (at the time of writing, Release 12).

3.2.2 Paging

There is a substantial amount of research work in the literature on paging mechanisms for wireless and cellular networks [49]–[54]. Some of this work has recently gained attention due to the irruption of novel deployment strategies for cellular networks, such as Heterogeneous Networks (HetNets) and small cells.

The authors in [49] propose a paging taxonomy based on the way the functional entities in the core network page the cells in the destination TA (i.e., cell flooding vs. selective paging strategies in order to reduce location signalling traffic throughout the network). These strategies comprise *sequential paging* [50], *selective paging* [51], *blanket paging* [52], *velocity paging* [53], and *pipeline paging* [54]. Below we provide a brief description of each one of these mechanisms.

In sequential paging schemes, users are gradually paged in subgroups of cells in the tracking area following a predefined order. These subgroups are commonly referred to as Paging Areas (PAs). The paging entity in the core network (e.g., the MME in the Evolved Packet System) sorts the PAs according to different parameters, such as UE mobility pattern, call arrival characteristics, overall signalling cost for each PA, paging delay minimisation, etc. A particular instantiation of sequential paging is selective paging. In this scheme, the paging entity in the core network sorts the PAs according to the probability of finding a user in each one of them. This probability is estimated from the call arrival and mobility pattern characteristics of each UE. In sequential (and, thus, selective) paging schemes, the number of cells in each PA must be kept under control in order to avoid long end-to-end delays during the execution of the paging procedure.

Blanket paging is the most common paging mechanism in cellular networks due to its simplicity and effectiveness. In blanket paging schemes, the mobility management entity in the core network sends paging messages to each cell in the destination TA upon reception of an incoming voice call or data connection request. When the destination UE receives a paging notification over the radio interface from its serving (H)eNB, it establishes an RRC connection with the core network. Blanket paging usually generates a large amount of location signalling traffic over the radio access network depending on the paging arrival rate and the number of cells in each TA. Consequently, this mechanism is not particularly efficient in large-scale networks of small cells.

Velocity paging schemes aim at capturing the individual UE's real-time mobility characteristics to provide an accurate prediction of the corresponding paging area. The most challenging issue in these mechanisms is to determine the actual speed of the UE with the necessary accuracy level. Since the destination UE in a paging procedure is in RRC idle mode, the device cannot report its position and/or speed values accurately to the mobility management entity in the core network. Velocity paging schemes usually characterise the mobility pattern of a given UE in terms of its velocity class or speed range (e.g., static, slow, medium, fast). Additionally, inaccuracies in paging area predictions lead to higher paging delays.

Pipeline paging schemes do not assume any previous knowledge of the location probabilities of the destination UE. Instead, the destination tracking area is divided into fixed PAs for all UEs, and multiple UEs are paged in a pipeline fashion. Incoming paging requests are queued in the core network paging entity, usually following a first-in, first-out (FIFO) discipline. Then, all PAs are paged sequentially until the destination UE initiates a Service Request procedure or, alternatively, all the PAs in the tracking area have been paged. The main difference between sequential and pipeline mechanisms is that, in the latter, several paging requests can be launched in a pipeline fashion during a single paging cycle. The main drawback of pipeline paging is that it must tolerate a high paging delay in order to achieve an optimal performance (due to the buffering of paging requests).

Instead of relying on simulation and/or modelling techniques to design and evaluate the performance of new paging mechanisms, some work in the literature takes a more empirical, data-driven approach [55], [56], [57]. For example, the authors in [55] processed more than 300 million anonymous call records from a CDMA2000[5] mobile network operator in the United States (Sprint) to characterise user mobility. In particular, they defined a set of mobility profiles based on parameters such as the location of a user during recent calls and call patterns in the previous days. Later on, they used these profiles to develop a family of paging techniques that adapt dynamically as user profiles get updated. The authors claim that their proposed schemes can reduce paging signalling traffic by 80% with a marginal increase in paging delay (approximately, less than 10%). In general, most data-driven paging schemes exploit historical mobility data to provide a more accurate estimation of user location and, thus, optimise the operation of selective paging algorithms.

As far as the standardisation work in 3GPP is concerned, the standard paging procedure for LTE networks has been defined in [58]. According to the taxonomy discussed in [49], this paging mechanism can be classified as blanket paging. To the best of our knowledge, no new 3GPP work items aimed at improving the paging procedure are expected to be included in the latest release of 3GPP Technical Specifications (at the time of writing, Release 12).

3.2.3 Conclusions

Standard 3GPP location management mechanisms have been designed with macrocell scenarios in mind. Thus, their performance in large-scale networks of small cells is far from optimal due to the large volume of location-related signalling generated by frequent handovers and cell reselections. This is particularly critical for all-wireless NoS, where the interconnection amongst small cells is realised via a wireless multi-hop backhaul. Thus, networks of small cells require customised paging and Tracking

[5] CDMA2000 is a family of 3G standards used mostly by some mobile network operators in the United States.

Area Update procedures in order to support efficient UE location services whilst keeping signalling traffic under control.

3.3 Traffic Management in Cellular Networks

As described in Section 2.4.1.1, the EPS bearer service is the standard unicast traffic management mechanism in the 3GPP Evolved Packet System. The purpose of the bearer service is to establish the necessary user-plane traffic flows (i.e., *bearers*) in order to enforce the corresponding QoS requirements across functional entities in the Evolved Packet System. As far as the transport network layer is concerned, these bearers are mapped onto GTP-U tunnels between adjacent functional entities in the RAN and the EPC. Overall, the EPS bearer service is a complex traffic management mechanism that involves a substantial amount of signalling traffic throughout the mobile operator's radio access and core networks.

The EPS bearer service architecture defines a traffic management framework where end-to-end user-plane data flows need to be established between a UE and the P-GW. This is particularly inefficient for small cell deployments (either NoS or standalone), as a UE might want to communicate with a peer entity in the same local network (e.g., another UE, a printer, a storage server, etc.) without the need to reach the P-GW in the core network. Similarly, a UE might want to establish a communication with an external peer entity that is directly accessible via its serving small cell's backhaul connection (e.g., an Internet server). As described in Section 2.4.1.5, Release 9 and 10 of 3GPP Technical Specifications introduced two modes of operation aimed at addressing local traffic issues in standalone small cell scenarios: LIPA and SIPTO. However, neither LIPA nor SIPTO support the establishment of direct user-plane bearers between UEs and/or local servers connected via the LTE-Uu interface to HeNBs in the same network of small cells (though not necessarily in the vicinity of each other). Thus, support for efficient local traffic in NoS (in terms of radio and core network resource usage) remains an open issue.

As far as multicast traffic is concerned, MBMS is the standard traffic management mechanism for supporting multicast/broadcast services in the 3GPP Evolved Packet System. The MBMS network and protocol architecture has been described in Section

2.4.2. Large-scale networks of small cells are particularly suitable to provide high-quality multicast/broadcast services in localised areas, such as sports venues, convention centres, or shopping malls. However, system throughput in MBMS is limited to the QoS achievable by users experiencing the worst radio conditions. This is due to the lack of QoS feedback mechanisms between UEs and the MBMS functional entities in the Evolved Packet Core (e.g., the MCE). To this extent, MBMS (and, by extension, multicast systems for wireless networks in general) underutilise radio resources in exchange for higher reliability at the user end.

Below we provide an overview of some of the most recent developments in unicast and multicast traffic management mechanisms for cellular and wireless networks. With regards to unicast traffic, we particularly focus on improvements to the EPS bearer service architecture for local traffic scenarios in networks of small cells, such as LIPA, SIPTO, and proximity-based services. As far as multicast traffic is concerned, we focus on some non 3GPP-specific QoS feedback mechanisms aimed at improving multicast system throughput in wireless (and cellular) networks.

3.3.1 Unicast traffic

LIPA, SIPTO, and proximity-based services are some of the most recent developments in unicast traffic management in cellular networks. Whilst LIPA and SIPTO focus on providing local IP access and selected traffic offload services in small cells scenarios, respectively, proximity-based services aim at supporting device-to-device (D2D) communications between close-by UEs (i.e., in the vicinity of each other). An overview of the state of the art for each one of these topics is provided below. In addition, references to both 3GPP Technical Specifications and research work in the literature have been included, whenever applicable.

3.3.1.1 LIPA and SIPTO Services for Small Cells

As described in Section 2.4.1.5, LIPA is a 3GPP Release 9 feature aimed at providing access for IP-capable UEs connected via LTE-Uu interface of a small cell to other IP-capable entities in the same residential/enterprise IP network, such as a local host, a storage server, a printer, etc. Similarly, SIPTO is a 3GPP Release 10 feature that provides a UE camped on a small cell with access to an external IP network (e.g. the Internet), without the need to traverse the MNO’s core network.

Most of the research work in the literature has traditionally focused on revisiting LIPA and SIPTO mechanisms for standalone small cell deployments. The authors in [60] discuss the different LIPA/SIPTO architectural alternatives introduced in [23], along with some related aspects such as network management, energy efficiency, radio enhancements, network capacity, and service continuity. Similarly, [61] provides a comprehensive tutorial on data offloading techniques for 3GPP Release 10 networks, with a particular focus on LIPA/SIPTO and IP Flow Mobility (IFOM). IFOM is a MAPIM (Multi-Access PDN Connectivity and IP Flow Mobility) 3GPP work item aimed at offloading traffic by moving user-plane data sessions between 3GPP and non-3GPP radio access networks (e.g., Wi-Fi). Whilst data offload in LIPA/SIPTO is transparent to the UE, offload logic in IFOM is mostly UE-centric, thus largely transparent to the RAN. On a different note, the authors in [62] propose a Universal LIPA service (uLIPA) aimed at providing a UE with access to its home network whilst being camped on a macrocell or visited LIPA network. Although their proposal is, in principle, only applicable to UMTS, the paper provides a comprehensive architectural and procedural framework that leverages the existing UMTS LIPA architecture and routing path establishment procedure by introducing the concept of local user-plane functional entities. In addition, it defines a new signalling procedure for mutual authentication between the UE and the visited (and home) LIPA networks.

As discussed in Section 3.1.1, the concept of networks of small cells was originally introduced in the context of the European Project BeFEMTO [4]. In this Ph.D. Thesis dissertation we particularly focus on all-wireless NoS as the prime reference scenario. Although LIPA provides connectivity to local IP entities in residential/enterprise small cell deployments (e.g., a storage server, a printer, etc.), there is a lack of support for the establishment of direct user-plane data bearers between UEs camped on the same network of small cells (though not necessarily in the vicinity of each other). Also, current local IP access solutions do not fully exploit the broad spectrum of possibilities of wireless multi-hop backhauls in the underlying transport network, such as an even distribution of radio and network resource usage throughout the entire NoS.

3.3.1.2 Proximity-Based Services for Small Cells

In addition to LIPA and SIPTO, device-to-device (D2D) communication scenarios have been recently proposed as a traffic offloading solution in cellular networks. To this extent, the Proximity-Based Services study item in Release 12 of 3GPP Technical Specifications (FS_ProSe) defines a new type of radio bearer aimed at supporting direct traffic exchange between a pair of close-by UEs (e.g., a few meters apart from each other) without involvement from user-plane functional entities in the EPC (i.e., S-GW, P-GW). This radio bearer is commonly referred to as the D2D bearer. Both direct communication between UEs (i.e., without (H)eNB involvement) and locally-routed (i.e., via the same serving (H)eNB) scenarios are supported by the D2D bearer. For a detailed description of applicable use cases, feature requirements, and charging and billing procedures for proximity-based services the reader is referred to [63], [64], and [65].

The authors in [66] build upon the 3GPP FS_ProSe study item and propose an architectural and procedural framework to support proximity-based services in the 3GPP Evolved Packet System. The aim of this proposal is to reduce EPC and radio congestion by localising data offloading points within UEs. In particular, they define a network architecture based on a new reference point for D2D traffic between close-by UEs (the Di interface), a set of D2D coordination functions localised in the P-GW, and a new D2D user-plane bearer to support direct traffic exchange between close-by UEs using a single link instead of an uplink and a downlink. In addition, the authors define a new Proximity Services Management function (PSM) in the NAS protocol, a new control-plane protocol stack architecture to support D2D enhancements in the UE and the MME, and a detailed message sequence chart for the D2D call establishment procedure. Similarly, the authors in [67] define an alternative architectural and procedural framework to support proximity-based services in the Evolved Packet System based on a new functional entity in the EPC called *the D2D Server*. Some of the functions of this entity comprise D2D-related device identifier allocation, policy management, UE-location assistance, call establishment, mobility tracking, etc. This raises the need of enhancements in some of the existing EPC functional entities (e.g., MME, HSS, PCRF) in order to support control-plane communication with the D2D server. Furthermore, the authors propose modifications to the NAS, RRC, RLC, and

MAC layers in the MME, eNB, and UE protocol stacks. In addition, they define a set of 3GPP network procedures aimed at enabling proximity-based services in the Evolved Packet System, such as D2D Service/Device Discovery and D2D Call Establishment and Management. Finally, the authors define the necessary message sequence charts to support both D2D-to-infrastructure and infrastructure-to-D2D user mobility scenarios. Alternative architectural and procedural frameworks for providing proximity-based services in the Evolved Packet System can be found in [68], [69], and [70].

On a different note, the authors in [71] and [72] advocate the use of out-of-band Wi-Fi Direct D2D technology to support proximity-based services between close-by UEs. The authors claim that the proposed mechanism can achieve higher data rates than those traditionally provided by the cellular infrastructure. This mechanism is realised by offloading D2D traffic onto high-capacity, unlicensed radio links (such as those in Wi-Fi Direct). Simulation results show capacity gains of up to 200% by offloading just 30% of user-plane data sessions onto Wi-Fi Direct D2D links. Furthermore, the authors claim that Wi-Fi Direct links can also provide energy efficiency gains of up to 4 times over very short distances. However, the authors fail to provide a description of the reference network scenarios and software tools that have been used to produce the simulation results.

It must be noted that the architectural and procedural solutions discussed above are not exclusively applicable to standalone small cells scenarios. Thus, proximity-based services can be also realised in large-scale networks of small cells by leveraging architectural frameworks borrowed from standalone deployments. However, there is a specific traffic scenario that is particularly suitable for all-wireless networks of small cells. As previously discussed, an all-wireless NoS consists of a large number of cooperative small cells that are interconnected via a wireless multi-hop backhaul. Thus, enabling direct user-plane traffic exchange between UEs camped on the same local network domain (but not necessarily in the proximity of each other) without the need for reaching an EPC user-plane entity might be particularly interesting from the point of view of network resource usage. Though this user-plane traffic scenario might not strictly be considered as a proximity-based service (as there is no direct link between the source and destination UEs), it most definitely falls within the scope of direct communications without involvement from EPC functional entities. To the best of our knowledge, there is no previous research work in the literature aimed at providing direct user-plane traffic exchange services between UEs in the context of large-scale, all-wireless networks of small cells.

3.3.2 Multicast traffic

As discussed in Section 2.4.2, MBMS is the standard 3GPP mechanism to provide multicast/broadcast services in the Evolved Packet System. Traffic management in MBMS relies on the use of SYNC in conjunction with IP multicast routing protocols between the MBMS GW and the participating eNBs (over the M1 interface), as well as on the Physical Multicast Channel (PMCH) in the LTE-Uu interface. In multicast and broadcast wireless transmissions, system throughput is usually limited to the QoS experienced by the users in the worst radio conditions (cell-edge users). In practice, this translates into the use of robust modulation and coding schemes (i.e., low bit rates) in order to guarantee a reliable reception of multicast/broadcast traffic for the maximum number of users in the coverage area.

As opposed to unicast traffic (where sender nodes perform bit rate adaptation based on feedback messages sent by the receivers), multicast traffic is essentially a one-way communication service. This is the case of MBMS, where UEs receive multicast traffic by decoding the Multicast Control and Multicast Traffic logical channels (MCCH and MTCH, respectively), but do not provide feedback to the sender eNB. In fact, strictly speaking, a UE participating in an MBMS session is in RRC idle mode (as there are no dedicated radio resources established with the network), unless it is simultaneously engaged in a unicast voice call or data connection. A common solution for improving multicast/broadcast services in wireless and cellular networks relies on introducing a small amount of feedback from selected receivers in order to enable rate adaptation in the sender node. This is commonly referred to as *feedback selection*. Since feedback selection techniques are, in general, RAT-agnostic, they can be applied to both wireless (e.g., Wi-Fi) and cellular (e.g., UMTS, LTE, LTE-A) scenarios. This subsection provides a general overview of recent developments in feedback selection techniques for wireless multicast scenarios.

The vast majority of solutions for improving multicast services in wireless and cellular networks rely on the integration of Automatic Repeat Request (ARQ) mechanisms into the protocol architecture [73]–[77], adding Forward Error Correction (FEC) packets to the multicast stream [78], [79]; or both [80]. Other studies propose rate adaptation mechanisms for improving network utilisation [81]. All previous examples rely on the provision of QoS feedback from end users to the sender node(s). Essentially, these feedback-gathering mechanisms can be classified into four main categories: *Individual Feedback from multicast receivers*, *Leader-Based Protocol with acknowledgements* (LBP-ACK), *Pseudo-Broadcast*, and *Leader-Based Protocol with negative acknowledgements* (LBP-NACK). Below we provide a brief description of each one of these schemes:

Individual Feedback mechanisms require all receivers to send acknowledgements of received packets. Using this feedback data, the sender node determines which packets have been lost and, in turn, proceeds to retransmit them to the multicast group. Although this approach guarantees a reliable distribution of multicast packets, it suffers from scalability issues due to the high volume of feedback overhead in large multicast/broadcast groups. The remaining approaches aim at reducing this overhead as follows.

The LBP-ACK approach [76]–[78], [80], [82], [83] solves the scalability issue by selecting a subset of receivers to provide feedback. The Pseudo-Broadcast approach [73], [74], [83], [84] converts the multicast feed into a unicast flow and sends it to one leader only, typically the receiver that is experiencing the worst channel conditions. Then, the leader acknowledges the reception of the unicast flow. The other nodes receive the multicast packets by listening to the channel in promiscuous mode. The LBP-NACK approach [75], [81], [85] improves the Pseudo-Broadcast mechanism by allowing other receivers to send NACKs for lost packets. After receiving the ACK from the leader, the sender can infer successful transmission to all receivers, as a NACK would collide with the leader's ACK.

As far as standardisation work in 3GPP is concerned, Release 10 of 3GPP Technical Specifications adopted MPEG DASH (Moving Picture Experts Group, Dynamic Adaptive Streaming over HTTP) [86] as the standard mechanism to encode multimedia content in MBMS. Essentially, MPEG DASH slices a multimedia stream into a sequence of independent media segments and then encodes each segment multiple times with a different set of quality parameters (e.g., low, medium, high quality). In a full-fledged MPEG DASH client-server implementation, media segments are delivered to end users as independent files via the HTTP protocol. This is done dynamically, i.e., users request media segments with different quality encoding settings through standard *HTTP Request* messages. This way, MPEG DASH clients adjust the quality of the multimedia stream to current available network resources. However, this is not the case for MBMS. Since MBMS is a unidirectional multicast service, a single quality representation for the entire multimedia stream is delivered to all UEs in the MBMS session via a standard multicast/broadcast file distribution mechanism. In fact, MBMS uses MPEG DASH exclusively for media encoding and encapsulation purposes, but not for the delivery of multicast streams.

In terms of QoS feedback, [87] defines an *MBMS Reception Reporting* procedure for MBMS receivers. The purpose of this procedure is to inform the BM-SC entity in the EPC about the outcome of a specific MBMS session. Thus, upon completion of an MBMS user service, a group of selected UEs may establish a communication with the BM-SC in order to report statistics about the MBMS service (e.g., packet loss, delay, jitter, achieved bit rate, etc.). As described in [87], the MBMS reception reporting procedure for download delivery scenarios (e.g., multicast file download) is mainly used to report either the reception of one or more files, statistics on the stream, or both. Instead, in the case of streaming delivery scenarios (e.g., multicast audio/video stream), the MBMS reception reporting procedure is mainly used to report statistics on the actual audio/video stream. Prior to the initiation of the MBMS reporting procedure, the BM-SC can determine the percentage of MBMS receivers that will provide reception reports about the MBMS session. These reports are sent as Extensible Markup Language (XML) messages via the HTTP protocol [87].

Although the MBMS reporting procedure allows UEs to send feedback messages to the BM-SC, it is not intended to provide live QoS measurements to perform dynamic radio and network resource adaptation during ongoing MBMS services (e.g., bit rate and MCS settings in MBMS eNBs). Instead, the MBMS reporting procedure is mainly initiated for statistical purposes after the completion of an MBMS session.

Thus, support for live QoS reporting from end users aimed at improving the overall quality of experience in MBMS services remains an open issue.

3.3.3 Conclusions

Support of efficient unicast and multicast traffic management in large-scale, all-wireless networks of small cells is a critical issue. In these scenarios, efficiency is generally measured in terms of radio and network resource usage over the wireless multi-hop backhaul. In the case of unicast traffic, support of direct multi-hop user-plane bearers between UEs camped on the NoS (though not necessarily in the vicinity of each other) is a key requirement for the deployment of distributed routing algorithms capable of optimising network resource usage in the underlying transport network. However, neither LIPA/SIPTO nor 3GPP proximity-based services provide the necessary architectural and procedural framework to support the establishment of such direct bearers between UEs.

As far as multicast traffic management is concerned, MBMS suffers from the traditional limitations of multicast/broadcast services in wireless networks, i.e., the lack of a feedback mechanism capable of improving the overall quality of experience for end users through the use of rate adaptation. Though some generic ARQ and FEC proposals for wireless networks have been discussed in the literature, support for a live QoS-feedback mechanism aimed at improving the overall quality of experience in MBMS services still remains an open issue.

Chapter 4

Research Question

In the previous chapters we have introduced the technological, research, and industrial background to build a general understanding of networks of small cells. First, we have defined the concept of NoS in the context of 3GPP cellular networks (Chapter 1). Secondly, we have provided a technical overview of the 3GPP Evolved Packet System, with a particular focus on the network model, the user- and control-plane protocol stack architectures, and some of the key network procedures (Chapter 2). Finally, we have discussed some relevant academic and industrial research work in three key research areas of large-scale deployments of networks of small cells, namely network architecture, location management (paging and Tracking Area Update procedures), and traffic management (Chapter 3).

In this chapter we formulate the fundamental question that constitutes the basis of the research work presented in this Ph.D. Thesis dissertation. In addition, we discuss the relevance of this research question in the academic and industrial communities, as well as its impact on cellular networks as a whole.

4.1 Question Formulation

This Ph.D. Thesis takes a system-level approach towards the realisation of an architectural and procedural framework to support efficient deployments of large-scale networks of small cells in the 3GPP Evolved Packet System. In particular, we make a case for all-wireless networks of small cells, where connectivity between HeNBs is provided through a wireless multi-hop backhaul. Thus, we formulate the fundamental research question in the following terms:

What is the architectural and procedural framework needed to support efficient traffic and mobility management mechanisms in massive deployments of all-wireless 3GPP Long-Term Evolution networks of small cells?

In order to provide a substantiated answer to this question, we define the following research objectives:

(a) Design a network architecture capable of supporting scalable deployments of all-wireless networks of small cells in the context of the 3GPP Evolved Packet System. This architecture, whenever possible, must minimise the impact on existing 3GPP protocol stacks, functional entities, and network procedures. This work is presented in Chapter 5.

(b) Design new paging and Tracking Area Update procedures to provide efficient location management services in large-scale deployments of all-wireless networks of femtocells. In these schemes, efficiency will be measured in terms of signalling traffic throughout the local NoS domain (i.e., the wireless multi-hop backhaul), as well as towards the MNO's Evolved Packet Core. This is described in detail in Chapter 6.

(c) Design efficient traffic management procedures to provide unicast and multicast services in large-scale networks of small cells. In particular, this comprises (a) new network procedures to support the establishment of direct bearers between UEs in the local NoS domain, (b) handover support for direct bearers, and (c) a lightweight feedback selection and rate adaptation mechanism to improve the quality of experience of the Multicast/Broadcast Multimedia Service (MBMS) in small cell deployments. This work is presented in Chapter 7.

4.2 Validity of the Question

In order to substantiate the validity of the fundamental research question we aim at addressing some of the shortcomings of previous research work in the areas of network architecture, location management, and traffic management. These limitations show up at various fronts. From the architectural point of view, novel approaches like Cloud RAN, RAN as a Service, and a Software-Defined Evolved Packet Core (Section 3.1.2) propose new architectures aimed at supporting the deployment of more efficient and cost-effective cellular networks. However, these proposals raise some concerns regarding (a) the operational expenditure (OPEX) associated with the ubiquitous provision of high-throughput, low-latency backhaul links with the Evolved Packet Core and (b) the impact on existing 3GPP Technical Specifications. Thus, despite having specific 3GPP functional entities to handle residential/enterprise deployments of standalone small cells (e.g., the HeNB GW, the SeGW), the expected increase of mobile traffic in large-scale deployments of NoS limits the scalability of 3GPP architectural solutions at both the control- and user-plane levels. As discussed in Section 3.1, the design of an efficient network architecture to support massive deployments of wired/wireless networks of small cells in the context of the 3GPP Evolved Packet System remains an open issue.

As far as location management is concerned, standard 3GPP mechanisms have traditionally been designed with macrocell scenarios in mind. This raises scalability, performance, and network congestion issues in the local network and the backhaul, particularly for all-wireless networks of small cells, where keeping the volume of control-plane traffic over the wireless multi-hop backhaul under control is critical. As discussed in Section 3.2, a substantial amount of research work on location management mechanisms has been proposed in the literature. However, most of these mechanisms are either non-3GPP compliant or based on unrealistic assumptions, which compromises their implementation in real-life 3GPP cellular networks. To this extent, we propose a new paging and Tracking Area Update procedures in order to support efficient location management in large-scale networks of small cells. In particular, these mechanisms take into account the specific constraints of all-wireless scenarios (e.g., limited radio and network resources), whilst minimising the impact on existing 3GPP functional entities, protocol stacks, and network procedures.

Finally, the issue of unicast and multicast traffic management in all-wireless networks of small cells is also addressed. On the one hand, though standard LIPA and SIPTO solutions provide users with local IP access and selected IP traffic offload services in standalone small cell deployments, there are no specific traffic management mechanisms aimed at supporting the establishment of direct unicast user-plane bearers[6] between UEs camped on the same NoS. To this extent, exploiting

---

[6] These bearers, though also referred to as *direct*, are different from the proximity-based services described in Section 3.3.1.2. In particular, they provide direct communication between UEs that might not necessarily be in the proximity of each other via a wireless multi-hop backhaul.

the broad spectrum of possibilities of wireless multi-hop backhauls in NoS (both at the mobile and transport network layers) remains an open issue. As far as multicast traffic management is concerned, MBMS services suffer from the traditional limitations of wireless multicast/broadcast traffic distribution mechanisms. As described in Section 3.3.2, in these scenarios system throughput is usually limited to the QoS that can be achieved by using robust modulation and coding schemes (i.e., those yielding low bit rates). To the best of our knowledge, there are no research proposals or 3GPP study/work items in the literature aimed at improving the quality of experience of MBMS users by leveraging lightweight feedback selection and rate adaptation mechanisms.

4.3 Impact on Cellular Systems

Networks of small cells are gaining attention as a promising solution for improving network coverage and boosting capacity in next-generation cellular networks. Though 3GPP Technical Specifications define a common standardisation framework for handset manufacturers, network vendors, and mobile network operators, there are still plenty of opportunities for improvement. In this subsection we discuss the impact of the solutions proposed in this Ph.D. Thesis on real-life cellular systems. In particular, we make a case for mobile network operators and network equipment vendors, as they are the industry players that will mostly benefit from 3GPP-compliant deployments of networks of small cells.

In terms of network architecture, the research work described in Chapter 5 introduces new 3GPP functional entities local to the network of small cells. The role of these entities is to keep control- and user-plane traffic within the NoS in order to alleviate the load on the backhaul link and, thus, prevent congestion in the mobile operator's Evolved Packet Core. In addition, some mobility and call establishment functions are also transferred to the NoS in order to address the scalability issues derived from massive deployments of networks of small cells. As far as implementation complexity is concerned, all new functional entities can be implemented through firmware updates to the RAN and EPC infrastructure, thus eliminating the need of purchasing new equipment or performing expensive hardware updates to existing infrastructure. In brief, MNOs can benefit from the BeFEMTO architecture by enabling large-scale deployments of networks of small cells whilst mitigating the impact on congestion, implementation, and scalability in their existing infrastructure.

With regards to location management, the paging and Tracking Area Update procedures described in Chapter 6 enable a more efficient use of the radio and network resources in a NoS without compromising key performance metrics such as signalling cost and paging delay. In terms of implementation complexity, the proposed mechanisms can be realised by firmware updates to the existing RAN and EPC infrastructure, thus reducing capital expenditure for MNOs. On the user side, no hardware or software changes are needed in mobile terminals. Furthermore, the use of standard 3GPP protocols, data structures (*Information Elements*), and network procedures benefits both MNOs and network vendors, as this reduces implementation complexity and mitigates the impact on future releases of 3GPP Technical Specifications.

Finally, the traffic management mechanisms described in Chapter 7 provide an efficient way of supporting unicast and multicast traffic services in networks of small cells. On the one hand, the concept of *direct bearers* (D-bearers, Section 7.1) enables the establishment of direct unicast user-plane bearers between UEs in the same NoS with minimal involvement from centralised functional entities. This is particularly convenient for all-wireless NoS scenarios, as it allows an even distribution of radio and network resource consumption throughout the wireless multi-hop backhaul. As far as multicast traffic is concerned, the proposed feedback selection and rate adaptation mechanism for MBMS (Section 7.2) enables MNOs to provide a better quality of experience to end users whilst minimising the impact on the existing infrastructure. In addition, the proposed mechanism leverages existing 3GPP protocols and network procedures, thus reducing both implementation complexity and standardisation impact on future releases of 3GPP Technical Specifications.

Chapter 5

An Architectural Framework for Networks of Small Cells

In this chapter we describe a 3GPP-compliant architecture to support large-scale deployments of networks of small cells in the Evolved Packet System. This architecture is commonly known as the *BeFEMTO EPS Architecture*, as it was developed in the framework of the European Project BeFEMTO [4]. Although our proposal does not make assumptions regarding the underlying transport network, we consider all-wireless networks of small cells as the prime reference scenario.

The standard 3GPP Evolved Packet System architecture has already been described in Section 2.1. Instead, in the following sections we will focus on the architectural and functional improvements needed to support large-scale deployments of networks of small cells in the context of 3GPP LTE cellular networks. In particular, we will present an evolution of the Release 10 3GPP EPS network architecture (i.e., *the BeFEMTO EPS Architecture*, Section 5.1), as well as the internal structure of the most relevant functional entities involved (Section 5.2). For the sake of completeness, the entire BeFEMTO Evolved Packet System Architecture has been presented in this chapter, thus effectively covering contributions from various BeFEMTO partners. This has been done to illustrate the impact of new traffic and location management mechanisms on the 3GPP Evolved Packet System architecture. Specific contributions regarding traffic and location management functions are discussed in Sections 5.2.1 and 5.2.2, respectively. A detailed description of the associated network procedures can be found in Chapter 6 and Chapter 7, respectively.

5.1 The BeFEMTO EPS Architecture

The aim of the BeFEMTO EPS Architecture is to confine user- and control-plane traffic within the NoS domain, hence offloading the backhaul link towards the

Evolved Packet Core. This is achieved by introducing a new functional entity in the EPS network architecture: the *Local Network of Small Cells Gateway* (LNGW). From the functional point of view, the LNGW acts as a network manager for local mobility, traffic management, access control, authentication, power management, and fault management in the edge of the NoS domain. In addition, the LNGW can also provide LIPA and SIPTO services. As far as network management is concerned, the LNGW allows the mobile network operator to manage the NoS as a single aggregate entity with respect to many of these functions, i.e., by hiding network internals to the EPC whenever possible whilst exposing the features the mobile operator needs to access.

Figure 25 shows the BeFEMTO Evolved Packet System Architecture, as defined in [88]. The functional entities and interfaces highlighted in orange have been either introduced or enhanced by the BeFEMTO EPS Architecture. In the case of functional entities, these comprise the LNGW, the HeNB, and the HeNB Management System (HeMS). As far as the network interfaces are concerned, the BeFEMTO EPS Architecture introduces the new S-rat (Remote Access Tunnel) and Type 1C' interfaces. These functional entities and reference points are described below.

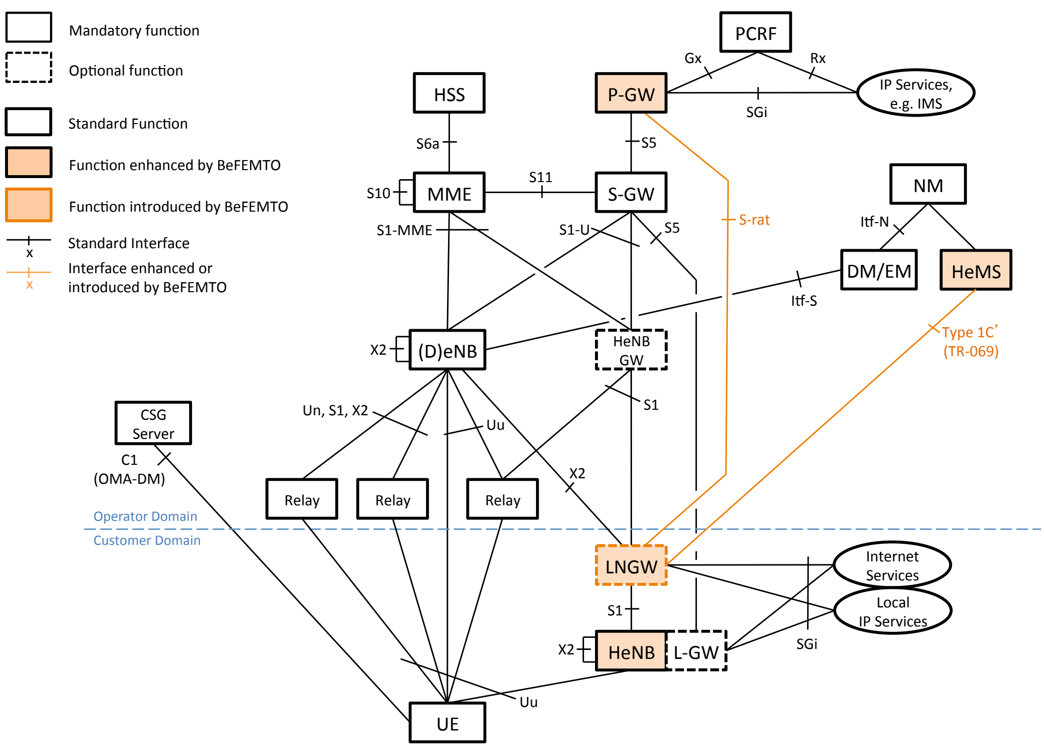


Figure 25.- The BeFEMTO Evolved Packet System Architecture

As shown in the figure, the BeFEMTO EPS Architecture defines two network domains, namely the *operator* and *customer* domains. The former comprises the functional entities deployed (and managed) by the mobile network operator; the latter encompasses the entities located within the scope of the local network operator, i.e., the organisational entity (not necessarily the MNO) in charge of deploying and managing the network of small cells. Below we provide a description of the functional entities introduced/enhanced by the BeFEMTO EPS Architecture in both the operator and customer domains:

- **Local Network of Small Cells Gateway (LNGW):** as previously mentioned, this functional entity constitutes the key building block of the BeFEMTO EPS Architecture. The LNGW acts as a concentrator for the S1-U and S1-MME interfaces, as well as a local mobility control entity for cell handovers. Additionally, it also acts as a HeNB controller for centralised radio resource and interference management support. Given the importance of the LNGW in the BeFEMTO EPS Architecture, a detailed description of its internal architecture can be found in Section 5.2.1.

- **Home evolved eNB (HeNB):** as described in Section 2.1.1, a HeNB is a customer-premises equipment that provides connectivity between a 3GPP UE and an LTE cellular network via the LTE-Uu interface. In addition, HeNBs are connected to the EPC of a mobile network operator via a broadband IP backhaul (e.g., ADSL, Ethernet, cable, optical fibre, etc.). The BeFEMTO EPS Architecture enhances standard 3GPP HeNBs by adding customised functions for Local Location Management and inter-HeNB routing, amongst others. These functionalities are described in Section 5.2.2.

- **HeNB Management System (HeMS):** the HeMS entity assumes the role of either an initial HeMS (optional) or of a serving HeMS. The initial HeMS may be used to perform identity and location verification of a HeNB and assigns the appropriate serving HeMS, security gateway and HeNB GW or MME to the HeNB. The serving HeMS supports HeNB identity verification and may also support HeNB GW or MME discovery. Typically, the serving HeMS is located inside the MNO secure network domain. Thus, the serving HeMS address is provided to the HeNB via the initial HeMS. In the BeFEMTO EPS

Architecture, the HeMS needs enhancements for supporting the LNGW, e.g., for providing the key settings that allow the LNGW to be inserted into the S1 interface, as well as local management policies.

- **Packet Data Network Gateway (P-GW):** as described in Section 2.1.1, the P-GW is the EPS functional entity responsible for connecting a UE to an external packet data network (e.g., the Internet). The P-GW acts as a user-plane mobility and inter-RAT handover anchor in the LTE cellular network. In addition, the P-GW is responsible for data bearer establishment, IP address allocation, Differentiated Services Code Point (DSCP) tagging, traffic filtering via traffic flow templates, and bit rate enforcement. As shown in Figure 25, the BeFEMTO EPS Architecture, the P-GW has been enhanced to handle the S-rat interface with the LNGW.

5.2 Functional Entities Supporting Networks of Small Cells

In this section we describe the architecture and functionalities of the most relevant entities introduced/impacted by the BeFEMTO EPS Architecture, namely the Local Network of Small Cells Gateway and the HeNB.

5.2.1 Local Network of Small Cells Gateway

In order to address the challenges that arise from large-scale deployments of networks of small cells, the BeFEMTO EPS Architecture has introduced the Local Network of Small Cells Gateway entity. As opposed to standard 3GPP HeNB gateways deployed by MNOs in the core network, LNGWs are deployed in the customer domain in order to act as a local controller for the entire network of small cells. Essentially, LNGWs aim at facilitating various functionalities for NoS deployments by acting as an interface between small cells and the Evolved Packet Core. Some of these functionalities comprise mobility management, routing, radio resource and interference management (RRIM), network operations and maintenance, and fault diagnosis for the local NoS domain. Essentially, LNGWs terminate control-plane signalling related to the customer domain, hence preventing it from traversing the backhaul link and reaching the MNO’s core network. In practice, this confines control- and user-plane traffic to the NoS scope, which results on lower processing load on EPC entities (in particular, on the HeNB GW and the MME). In addition, the local network manager function can enforce its own locally optimised policies, which reduces the volume of backhaul and management traffic towards the EPC.

Figure 26 shows the internal architecture of the LNGW functional entity. As shown in the figure, the LNGW can be divided into two main functional blocks. The upper block is related to management and fault diagnosis functionalities performed at a higher level. The lower block deals with the radio and network functionalities provided by the LNGW. These blocks are supported by several functional entities, as shown in the figure. In particular, the Proxy MME (P-MME) and Proxy S-GW (P-SGW) act as the local counterpart of the MME and S-GW entities in the Evolved Packet Core for supporting local mobility management and data forwarding functions in the network of small cells. Note that both the P-MME and the P-SGW appear to the EPC as a single HeNB node. Consequently, these entities can use standard 3GPP interfaces and procedures to provide traffic and mobility management functions.

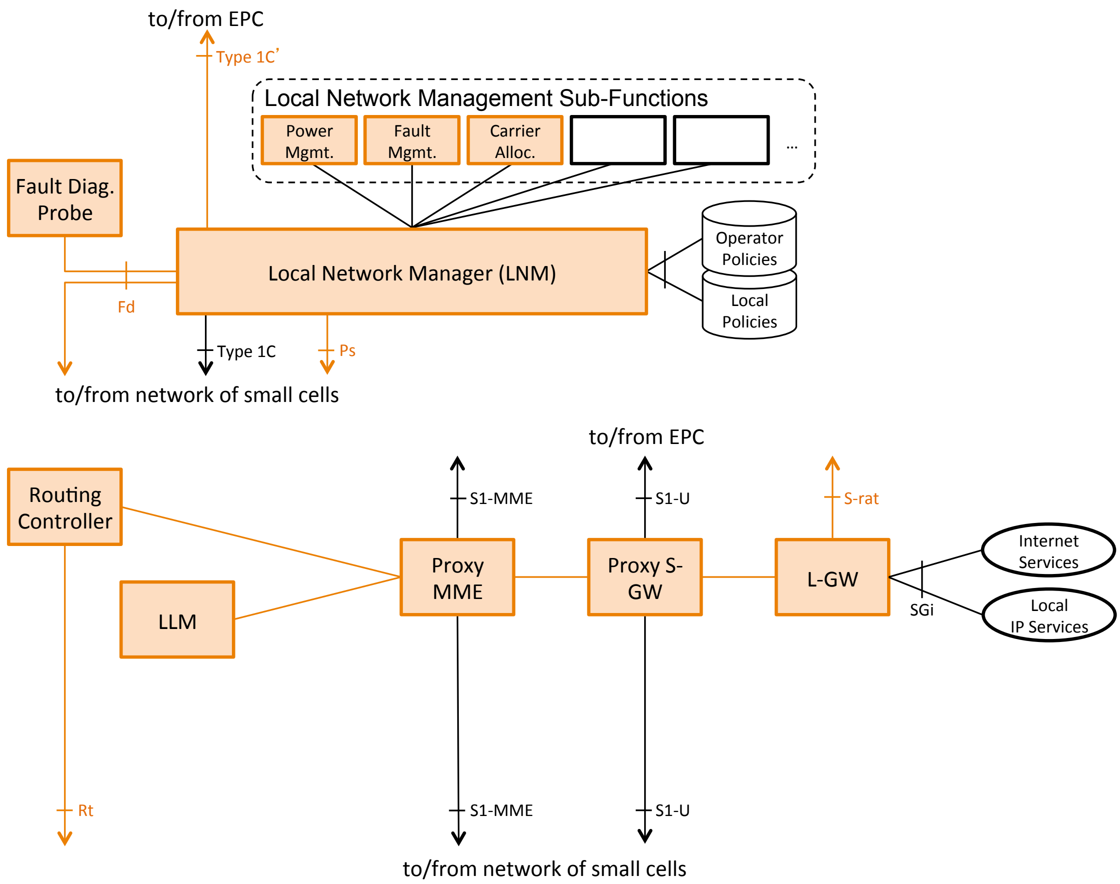


Figure 26.- The BeFEMTO LNGW functional architecture

In addition to the P-MME and the P-SGW, the LNGW also contains a Local P-GW functional entity (L-GW) to provide a local breakout point with full mobility support for LIPA and SIPTO sessions. By adding a new interface between the L-GW and the P-GW in the Evolved Packet Core (i.e., S-rat), LIPA/SIPTO mobility is extended to/from outside of the network of small cells. With regards to mobility management, the Local Location Management function processes standard 3GPP messages from the P-MME to support paging and Tracking Area Update procedures in the network of small cells. This is explained in detail in Chapter 6. As far as traffic management is concerned, the routing function enables local routing mechanisms to optimise user-plane traffic forwarding within the NoS (see Chapter 7). Additional functionalities comprise RRIM (responsible for coordinating the radio access in each small cell) and network management (supported by a Local Network Manager (LNM) based on local information collected via network probes).

5.2.1.1 Functional Blocks

The LNGW comprises the following radio and network functions:

- **Proxy Mobility Management Entity (P-MME):** as previously mentioned, the P-MME functional entity handles local mobility-related signalling functionalities by acting as a termination point for control-plane traffic from the NoS. The P-MME appears to the MME (or HeNB GW, if present) in the EPC as a HeNB, and to the HeNB in the NoS as an MME. Thus, its presence does not impact existing 3GPP network procedures in the UE and the MME.

- **Proxy Serving Gateway (P-SGW):** the P-SGW entity provides local routing and forwarding user-plane traffic services in the network of small cells. Thus, S1 bearers established by the Evolved Packet Core are terminated at the P-SGW and, in turn, mapped onto local S1 bearers towards small cells in the customer domain. In addition, the P-SGW acts as a local mobility anchor for inter-HeNB handover by conveying user-plane data from the S-GW in the core network. Analogously to the P-MME, the P-SGW appears to the S-GW (or HeNB GW, if present) in the EPC as a HeNB, and as a S-GW to HeNBs in the network of small cells. Thus, its presence in the BeFEMTO EPS Architecture does not impact existing 3GPP network procedures in the UE and the S-GW.

- **Local Gateway (L-GW):** the L-GW functional entity acts as a local P-GW connecting the NoS to external packet data networks. The L-GW has been enhanced with the functionality for consistent macro-to-small cell offload mobility, as well as remote access. Depending on the specific architecture, the L-GW can constitute a central local breakout point for LIPA and SIPTO traffic. However, as described in Section 7.1, direct user-plane traffic between UEs in the NoS can also be supported without L-GW involvement.

- **Routing Controller (RC):** the routing controller entity has different functionalities depending on the underlying routing scheme in the transport network layer. In the particular case of all-wireless networks of small cells, we assume a distributed routing scheme that exploits location information to forward traffic between HeNBs and from/to LNGWs. Further information on this routing scheme can be found in [89]. At the functional level, the routing controller interacts with the P-MME to obtain information about the established user-plane bearers and their corresponding QoS requirements.

- **Local Location Management:** the LLM function maps a given mobile subscriber identifier to the serving HeNB where the UE is currently camped on. Essentially, the LLM function interacts with the P-MME to handle paging requests from the Evolved Packet Core and find the corresponding UE in the network of small cells. A detailed description of location management procedures in large-scale networks of small cells can be found in Chapter 6.

In addition to the functional entities described above, the LNGW also comprises the following management functions:

- **Fault Diagnosis Probe (FDP):** this function provides fault diagnosis information related to problems in the LNGW to the Local Network Manager. Together with the information provided by FDPs deployed in HeNBs, this will be the basis for the diagnostic inference carried out by the Fault Management module.

- **Local Network Manager:** this function offloads management complexity from the MNO onto the LNGW in order to improve the scalability of networks of small cells. In particular, it enforces mandatory network management policies from the MNO in conjunction with local operational policies defined by the LNO. The Local Network Manager entity hosts various local network management components, such as the Power Management, Fault Management, and Frequency Allocation modules. Additional management components can be deployed to extend the functionalities of the LNM.

- **Power Management (PM):** the PM function tracks the UEs' activities in the local network and adaptively enables the unused small cells into power-saving mode to reduce the overall power consumption in the NoS. Depending on the specific implementation scenario, various monitoring methods and power-saving modes can be adopted (e.g., a self-organising mechanisms to enable UEs to learn on their own when they are in the geographical vicinity of the NoS). Further information on power management schemes in the context of the BeFEMTO Project can be found in [90].

- **Fault Management (FM):** this function is the core fault diagnosis component in the LNGW architecture. The FM function uses information provided by FDPs in HeNBs and the LNGW to determine the root cause of network issues. FDPs may need to communicate with other FDPs located in the mobile and access network providers to execute distributed inference. Fault management mechanisms in NoS are beyond the scope of this Ph.D. Thesis dissertation. For further information the reader is referred to [91].

- **Carrier Allocation (CA):** the CA function in the LNGW enables the local network operator to assign LTE carriers to HeNBs according to their traffic demands and QoS policies. These local policies must be compliant with high-level policies determined by the mobile network operator. For further information about carrier allocation schemes in networks of small cells the reader is referred to [92].

5.2.1.2 Network Interfaces

As shown in Figure 26, the LNGW interacts with other functional entities in the BeFEMTO EPS Architecture through the following network interfaces:

- **S1-MME:** the S1-MME interface towards the EPC is the reference point through which the LNGW exchanges standard mobility- and session-related signalling messages with the MME. Analogously, the LNGW uses the S1-MME interface towards the network of small cells to convey session/mobility messages to the corresponding HeNBs.
- **S1-U:** the S1-U reference point conveys user-plane data traffic between the P-SGW and the S-GW, and the P-SGW and HeNBs, respectively.
- **SGi:** this interface is used to exchange LIPA and SIPTO traffic between the LNGW and the NoS, as well as between the LNGW and the Internet. Local traffic within the NoS (e.g., D-bearers) can also be established without the involvement of the L-GW, as described in Chapter 7.
- **S-rat:** the LNGW uses the S-rat interface to communicate with the P-GW in the Evolved Packet Core. This interface provides service continuity for LIPA and SIPTO sessions during a handover from a NoS to the macrocell network. Additionally, the S-rat manages the re-establishment of such sessions whilst the UE is camped on an eNB.
- **Ps:** the Ps interface is used to convey measurement requests from the LNGW, as well as measurements results from the addressed HeNBs in the network of small cells.
- **Rt:** the Routing Controller in the LNGW uses the Rt interface to obtain link state information, update events, and modify the corresponding forwarding tables in the underlying transport network layer.
- **Fd:** the FDP function uses the Fd interface to send and receive fault diagnosis information to/from the Fault Management module located in the LNGW and other HeNBs in the network of small cells.
- **Type 1C':** this optional interface between the HeMS and the LNGW is used to provide configuration, software updates, network monitoring and other management functionalities for the LNGW and the NoS as a whole.

5.2.2 Modifications to HeNBs

3GPP Technical Specifications define the architectural and procedural framework that different cellular industry players (handset manufacturers, network vendors, mobile network operators) must observe. Nevertheless, handset manufacturers and network vendors are free to include proprietary modules in their products in order to introduce new functionalities or extend the existing ones. However, this approach is subject to interoperability constraints in order to ensure a seamless operation amongst products from different manufacturers. The BeFEMTO HeNB Architecture builds upon the idea of extending the standard 3GPP framework with new and enhanced functional blocks to address specific issues derived from large-scale deployments of all-wireless networks of small cells. The underlying assumption is that new and extended functionalities pose a minimal (or, sometimes, non-existent) impact on current 3GPP Technical Specifications. Figure 27 describes the BeFEMTO HeNB Architecture, as defined in [88].

As shown in the figure, the BeFEMTO HeNB Architecture comprises a large variety of network protocols and functions, ranging from Self-Organising Network (SON) enablers to traditional control- and user-plane protocol stacks, amongst other. These functional blocks may or may not be implemented in commercial network equipment depending on the specific deployment scenario (e.g., standalone vs. networks of small cells). As far as the BeFEMTO HeNB Architecture is concerned, enhanced or newly introduced functional blocks have been highlighted in orange. Amongst other, these comprise management functions, control- and user-plane protocol stacks, physical layer procedures, SON enablers, network functionalities, etc. Enhanced or newly introduced functional blocks specifically related to traffic and mobility management procedures have been highlighted in red.

The middle section in Figure 27 shows the user- and control-plane protocol stacks for both the wireless backhaul interface (middle-left) and the LTE-Uu interface (middle-right). Each layer in the protocol stack architecture contains the necessary state machines and network mechanisms to ensure a basic operation of the protocol. Instead, the control logic associated to each protocol is located in a separate functional block in order to facilitate algorithm replacement and cross-layer/joint operation. For example, the MAC layer in the LTE-Uu protocol stack (middle-right section in Figure 27) implements the standard MAC multiplexing mechanisms in LTE. However, the MAC scheduling logic has been factored out into a separate scheduler module. Similarly, the Self-Organising RRIM functions of the RRC protocol in the LTE-Uu interface have been implemented in an external SO-RRIM module.

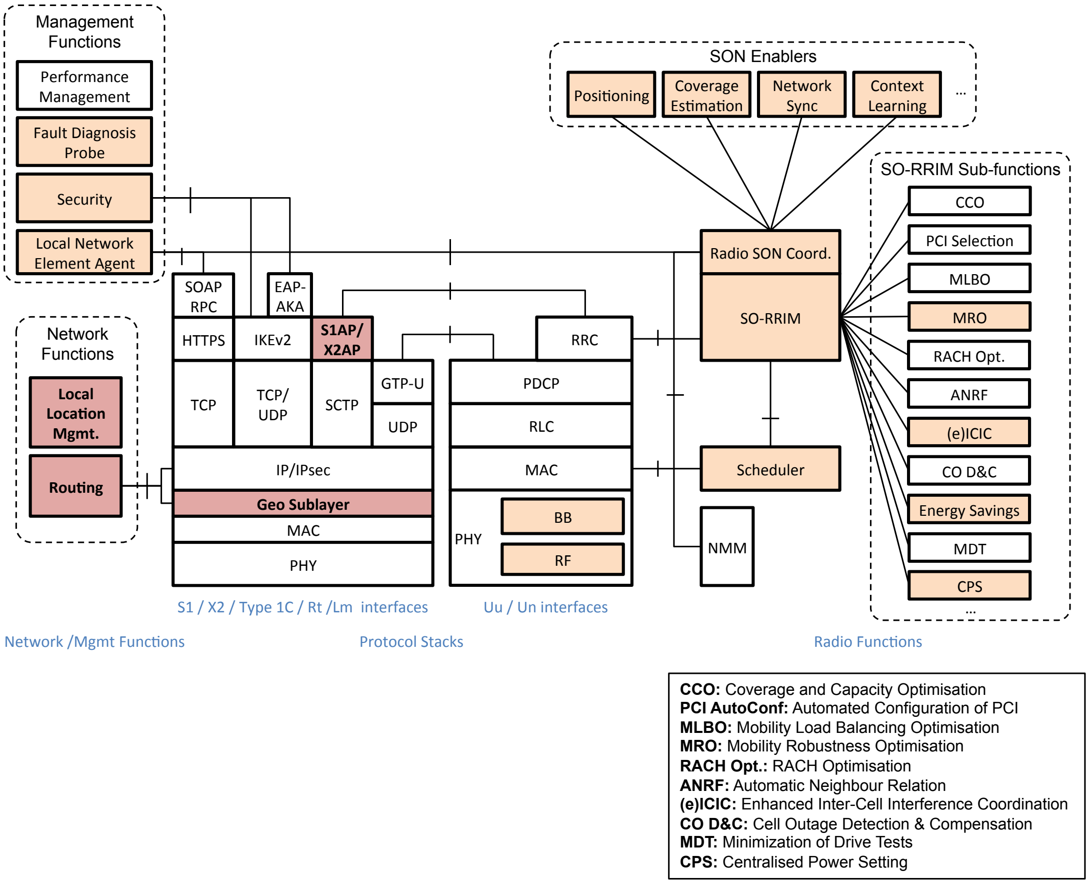


Figure 27.- The BeFEMTO HeNB Architecture

The rightmost section in Figure 27 shows the functional blocks related to internal radio functions in the BeFEMTO HeNB architecture. These comprise radio resource management, self-configuration, and self-optimisation of radio parameters. The Radio SON Coordinator plays a central role, as it combines the information provided by the SON-enabling functions with operator policies and parameters received from Operations, Administration, and Management entities (OA&M) in order to achieve the required performance goals. These parameters are related to the Self-Optimising Radio Resource and Interference Management functional block (SO-RRIM). The SO-RRIM functional entity optimises radio resource usage within and between HeNBs with the assistance of various SO-RRIM strategies. In turn, the lower level scheduler performs resource allocation between radio bearers subject to the constraints imposed by the SO-RRIM module, and potentially providing feedback to the SO-RRIM whenever needed. Finally, the leftmost section in Figure 27 depicts the network-related functional blocks in the BeFEMTO HeNB Architecture. The upper group is related to network management functions in the HeNB (e.g., security and performance management). On the other hand, the lower group is related to operations taking place in the underlying transport network layer, such as user- and control-plane traffic routing via the wireless backhaul interface.

A comprehensive description of all the functional entities involved in the BeFEMTO HeNB Architecture is provided in [88]. For the sake of simplicity (and due to space constraints), this section focuses on the functional blocks involved in traffic and mobility management procedures only. These blocks are described below:

- **Local Location Management:** the local location management function in the BeFEMTO HeNB Architecture manages the necessary 3GPP mechanisms to track the location of UEs in the network of small cells. As described in Section 2.3, user location in the Evolved Packet System is supported by the paging and Tracking Area Update mechanisms. The BeFEMTO EPS Architecture introduces new paging and Tracking Area Update procedures aimed at improving the efficiency of standard location management schemes in all-wireless networks of small cells. These mechanisms are described in detail in Chapter 6.
- **Routing:** the routing block supports control- and user-plane traffic routing between HeNBs and the LNGW. As opposed to standalone small cell deployments, where HeNBs communicate with UEs via the LTE-Uu interface and with EPC entities via the backhaul link, HeNBs in all-wireless networks of small cells have the ability to route traffic over the wireless multi-hop backhaul in a distributed fashion. Amongst other requirements, this involves finding an optimal route between the source and destination HeNBs in terms of available radio and network resources. Chapter 7 describes the necessary 3GPP procedures to establish direct user-plane bearers (D-bearers) between UEs camped on the same network of small cells. For a detailed description of

the distributed routing protocol used in the underlying transport network layer the reader is referred to [89].

- **Geo Sublayer:** the geographic sublayer (Geo Sublayer) in the wireless backhaul protocol stack is the key component of the distributed routing protocol described in [89]. Essentially, this protocol exploits geographic coordinates (either physical or virtual) of HeNBs in order to support efficient traffic routing in an all-wireless network of small cells. In this context, the Geo Sublayer function obtains the necessary geographic information (e.g., HeNB physical/virtual coordinates) to perform protocol encapsulation in the form of the so-called *geopacket*. This geopacket is, in turn, passed onto the MAC layer of the wireless backhaul interface to reach the destination node through the wireless multi-hop backhaul. A detailed description of the role and functions of the Geo Sublayer is provided in [89] and [93].

- **S1-AP:** the S1-AP protocol in the BeFEMTO HeNB Architecture has been modified to accommodate a distributed paging mechanism for large-scale, all-wireless networks of small cells. In particular, we have defined a new Information Element in the S1-AP protocol specification to encapsulate a bijective paging tree in the form of a sequence of HeNB IDs. The distributed paging mechanism and its implications in the EPS architecture are discussed in detail in Section 6.2.

- **X2-AP:** similarly to S1-AP, the X2-AP protocol has also been modified to accommodate the distributed paging mechanism for all-wireless networks of small cells. In particular, we have introduced a new Information Element in the X2-AP protocol specification to encapsulate the distributed paging tree, as well as a new X2-AP message (*X2-AP Paging*) to exchange paging messages between peer X2-AP protocol entities in neighbouring HeNBs. The impact of the distributed paging mechanism on the X2-AP protocol is described in detail in Section 6.2.

5.3 Research Contributions

In the course of the design of the BeFEMTO Evolved Packet System Architecture, the following contributions to the literature have been made:

- J. Ferragut, J. Mangues-Bafalluy, J. Núñez, F. Zdarsky, "Traffic and Mobility Management in Networks of Femtocells", ACM/Springer Mobile Networks and Applications Journal, Special Issue on Cooperative and Networked Femtocells, Vol. 2012, pp. 662-673, August 2012. Impact factor (2012): 1.109
- J. Núñez, J. Ferragut, J. Mangues-Bafalluy, "On Stateless Routing for an All-wireless Network of Femtocells. Implications in the 3GPP Architecture", in Proc. 73rd IEEE Vehicular Technology Conference (VTC 2011 - Spring), Workshop on Broadband Femtocell Technologies, 15-18 May, 2011, Budapest (Hungary). Acceptance rate (2011): 51.2%
- INFSO-ICT-248523 BeFEMTO – Deliverable D2.2, v1.0, "The BeFEMTO System Architecture" [Online]. Available: http://www.ict-befemto.eu
- INFSO-ICT-248523 BeFEMTO – Deliverable D2.3, v1.0, "The BeFEMTO System Concept and its Performance" [Online]. Available: http://www.ict-befemto.eu
- INFSO-ICT-248523 BeFEMTO – Deliverable D5.1, v1.0, "Femtocell access control, networking, mobility, and management concepts" [Online]. Available: http://www.ict-befemto.eu
- INFSO-ICT-248523 BeFEMTO – Deliverable D5.2, v1.0, "Femtocell access control, networking, mobility and management mechanisms (final)" [Online]. Available: http://www.ict-befemto.eu
- INFSO-ICT-248523 BeFEMTO – Deliverable D5.3, v1.0, "Evaluation report of femtocells networking, mobility and management solutions" [Online]. Available: http://www.ict-befemto.eu
- INFSO-ICT-248523 BeFEMTO – Deliverable D6.1, v1.0, "Selection of scenarios for proof of concept testbeds and specifications for key building blocks functionalities and interfaces" [Online]. Available: http://www.ict-befemto.eu

Chapter 6

Location Management Mechanisms in Networks of Small Cells

The large-scale deployment of networks of small cells in the 3GPP Evolved Packet System has implications on the paging and Tracking Area Update procedures. These limitations show up at various fronts. First, standard location management mechanisms have been designed in the context of macrocell networks, where handovers and cell reselections are less frequent than in NoS. Second, these mechanisms often rely on overprovisioned backhaul links between eNBs and the EPC, a common practice in macrocell scenarios that does not necessarily apply to large-scale, all-wireless NoS. Whilst network deployment strategies (macrocell- vs. NoS-based) have little or no impact on user mobility, UEs camped on networks of small cells are more prone to perform a higher number of handovers and cell reselections (and, in principle, more Tracking Area Updates) due to the smaller cell size. This additional volume of TAU-related signalling puts a strain on the limited RAN and backhaul resources of an all-wireless NoS. As far as the paging procedure is concerned, the standard 3GPP paging mechanism is particularly inefficient in NoS scenarios. 3GPP Technical Specifications require the MME to send a paging message to each one of the (H)eNBs in the UE's current tracking area, thus generating a substantial amount of paging-related signalling from the EPC to the RAN. This raises scalability and congestion issues in the wireless multi-hop backhaul.

The best-case scenario for location management in cellular networks is the one where signalling for *both* paging and Tracking Area Update procedures can be minimised simultaneously. In practice, this involves deploying a fine-grained user tracking mechanism that can determine the location of a UE with single-cell granularity without the penalty of triggering a Tracking Area Update every time the UE enters a new cell. However, this ideal case clashes with the underlying trade-off between the paging and Tracking Area Update concepts. In current 3GPP standards, minimising TAU-related signalling means registering UEs to bigger tracking areas, so that TAU procedures need to be executed less often. This necessarily comes at the expense of a more inefficient paging procedure, as the Evolved Packet Core will have to page more cells whenever a UE needs to be located. Analogously, in an optimal paging scenario, the EPC would only need to page the UE's current serving cell. However, this comes at the expense of a very inefficient Tracking Area Update procedure, as UEs would have to update their location every time they reselect or hand over to a new cell. The tight relation between the paging and Tracking Area Update procedures stresses the idea that both mechanisms are simply two sides of the same coin. Therefore, breaking the coupling between them is a key step towards achieving more efficient location management mechanisms in cellular networks.

In this chapter we propose new paging and Tracking Are Update procedures for large-scale deployments of all-wireless network of small cells. These mechanisms aim at reducing the amount of location-related signalling over the wireless multi-hop backhaul whilst mitigating the impact on existing 3GPP Technical Specifications. Section 6.1 discusses the design and evaluation of a self-organising Tracking Area List (TAL) mechanism. A distributed paging mechanism for all-wireless networks of small cells is described in Section 6.2.

6.1 A Self-Organising Tracking Area List Mechanism

As previously mentioned, standard 3GPP location management mechanisms have traditionally been designed with macrocell scenarios in mind. Therefore, their performance in large-scale, all-wireless networks of small cells is far from optimal due to the location-related signalling generated by frequent handovers and cell reselections. Thus, NoS require customised location management solutions in order to track mobile terminals efficiently whilst keeping signalling traffic under control.

In this section we propose a self-organising Tracking Area List mechanism aimed at reducing location management traffic in all-wireless networks of small cells. Although SON use cases have been extensively discussed in 3GPP [94], the focus when dealing with mobility management is mainly on handover management instead of on exploiting the benefits of Tracking Area Lists. As described in Section 3.2.1, TALs are a feature first introduced in Release 8 of 3GPP Technical Specifications aimed at allowing the MME to register UEs to more than one Tracking Area (TA) simultaneously. Our proposal leverages this 3GPP feature to reduce location-related signalling in large-scale networks of femtocells. This is achieved by registering UEs to individual TALs capable of adjusting their size in a self-organising fashion, i.e., without the need for common and static pre-defined network planning. Analytical results show that our self-organising TAL mechanism can generate up to 39% less location signalling traffic per UE than a conventional (i.e., static) TAL mechanism. In terms of implementation complexity, our scheme can be easily deployed in P-MMEs by means of a software update. In addition, it does not require modifications to the existing protocol stacks and 3GPP network procedures.

In the following subsections we discuss the operation of the self-organising Tracking Area Update mechanism from an analytical point of view. First, Section 6.1.1 introduces a mobility model aimed at characterising the mobility of UEs in dense cellular deployments. Secondly, Section 6.1.2 defines the network procedures upon which the self-organising Tracking Area List mechanism has been built. Finally, Section 6.1.3 combines the mobility model and mechanism description to carry out a performance evaluation of the proposed self-organising TAL procedure in various network scenarios.

6.1.1 Mobility Model

The author in [95] provides a comprehensive survey of analytical mobility models for cellular networks. In order to facilitate the mathematical tractability of the self-organising TAL mechanism, we have assumed a Markov-based mobility model for a 2D hexagonal network topology. This mobility model is described below.

The concept of slotted time is implicit to all Markov-based mobility models. Thus, at the end of a given timeslot, the UE may remain in the current cell with probability p or, alternatively, may transit to an adjacent cell with probability $\frac{1-p}{6}$. Variations in the UE speed can be modelled by modifying the value of p (i.e., values of p close to 1 are associated with slow/static UEs, whilst values of p close to 0 correspond to fast-moving UEs).

Let us consider that the cell topology depicted in Figure 28 corresponds to a given tracking area layout. We define the concept of *cell ring* as the group of adjacent cells where a UE can be found in a certain timeslot. This way, the first ring is formed by a single cell located in the centre of the TA (the white hexagon in Figure 28). Analogously, the second ring is formed by all one-hop external neighbours surrounding the first ring (the blue hexagons). The third ring is formed by all one-hop external neighbours encircling the second ring (the green hexagons), and so on. Following this reasoning, several concentric cell rings constitute a tracking area.

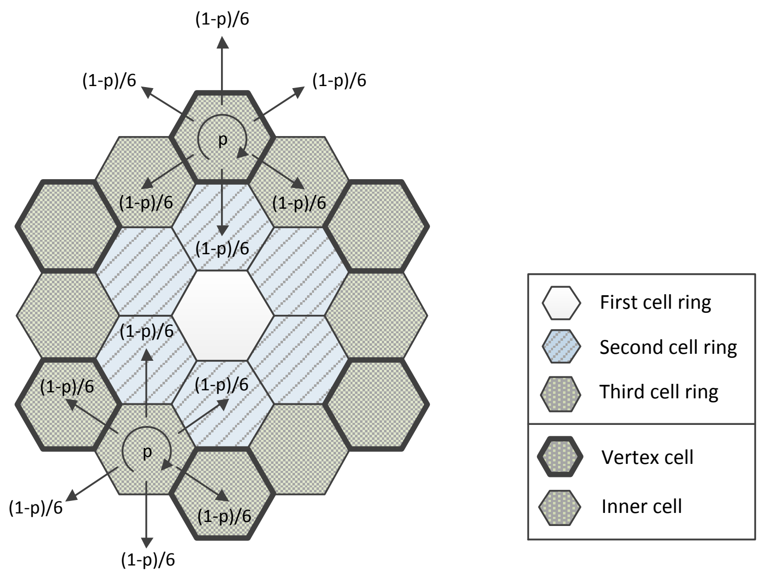


Figure 28.- Markov-based mobility model in a 2D hexagonal cell topology

The concept of *inner* and *vertex* cells is also depicted in Figure 28. By definition, both vertex and inner cells belong to the outermost ring of a tracking area. A vertex cell is a cell in the outermost ring of a tracking area that provides three exit points from the TA (vertex cells are identified by thicker hexagons in Figure 28). Analogously, an inner cell is a cell in the outermost ring of a tracking area that provides two exit points from the TA (i.e., the thin hexagons in the green cell ring). The difference between vertex and inner cells will be considered later during the computation of the average Tracking Area Update requests arrival rate, $\bar{\lambda}_{TAU}$.

Markov-based mobility models can be mathematically described by discrete-time Markov chains (DTMCs), as shown in Figure 29. This figure depicts the DTMC of a Markov-based mobility model in a hexagonal cell topology formed by N cell rings.

States in the DTMC represent different cell rings in the tracking area. Transitions between states represent mobility events in the TA, i.e., a UE staying in the current cell ring or, alternatively, moving (outwards or inwards) to an adjacent one. As shown in the figure, each state is labelled with a number ranging from 1 (innermost TA ring) to N (outermost TA ring). Analogously, transitions between states have been labelled with different probability values, namely $p\{stay, i\}$, $p\{next, i\}$, and $p\{back, i\}$.

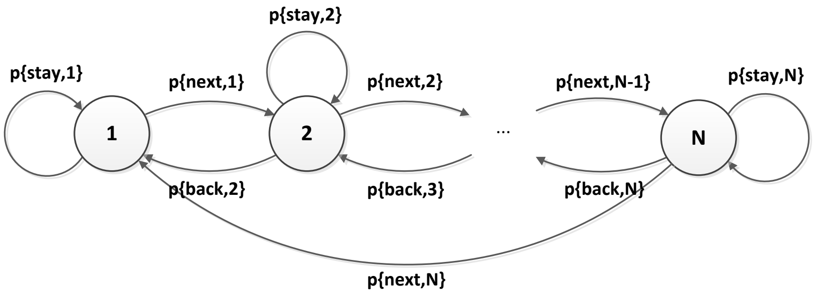


Figure 29.- DTMC of a Markov-based mobility model with N cell rings

As shown in the figure, $p\{stay, i\}$, $p\{next, i\}$, and $p\{back, i\}$ correspond to the probabilities of *staying at*, *moving outwards from*, and *moving inwards from* ring *i*, respectively. According to the cell topology depicted in Figure 28, these probabilities can be expressed as follows:

$$p\{stay, i\} = p + 2p'$$

$$p\{next, i\} = p\{vertex, i\} \cdot 3p' + p\{inner, i\} \cdot 2p'$$

$$p\{back, i\} = p\{vertex, i\} \cdot p' + p\{inner, i\} \cdot 2p'$$

Where:

$$p' = p\{UE\ crossing\ a\ cell\ boundary\} = \frac{1-p}{6}$$

$$p\{vertex, i\} = \frac{1}{i-1}$$

$$p\{inner, i\} = 1 - p\{vertex, i\} = \frac{i-2}{i-1}$$

Once all transition probabilities have been determined (i.e., the values of $p\{stay, i\}$, $p\{next, i\}$, and $p\{back, i\}$ have been computed $\forall i$ s.t. $1 \leq i \leq N$) we can calculate the one-step transition probability matrix of the DTMC ($\boldsymbol{P_{mob}}$). Essentially, $\boldsymbol{P_{mob}}$ captures the dynamics of the mobility model, as it contains all the transition probabilities between DTMC states. Later on, we use $\boldsymbol{P_{mob}}$ to calculate the steady-state probability vector of the DTMC ($\boldsymbol{\pi_{mob}}$), i.e., the vector that contains the probabilities of finding a UE in each one of the cell rings in the TA. Thus, by applying DTMC theory we can prove that:

$$\boldsymbol{\pi_{mob}} = [\pi_{mob}(1)\ \dots\ \pi_{mob}(N)] = \boldsymbol{e} \cdot (\boldsymbol{P_{mob}} + \boldsymbol{E} - \boldsymbol{I})^{-1}$$

Where $\pi_{mob}(i)$ is the probability of finding a UE in the *i*-th cell ring of the tracking area, $\boldsymbol{e}$ is a row vector of all ones, $\boldsymbol{E}$ is a matrix of all ones, and $\boldsymbol{I}$ is the identity matrix.

Once $\boldsymbol{\pi_{mob}}$ is known, we can calculate the probability of initiating a Tracking Area Update procedure in a TA formed by N cell rings as follows:

$$p\{TAU, N\} = \pi_{mob}(N) \cdot p\{next, N\}$$

Where $\pi_{mob}(N)$ is the probability of finding a UE in the outermost ring of a tracking area formed by N cell rings and $p\{next, N\}$ is the probability of the UE moving out of that ring.

Once $p\{TAU, N\}$ is known, we can calculate the average TAU request arrival rate generated by a UE camped on a tracking area formed by N cell rings, $\bar{\lambda}_{TAU,N}$. In order to do so, let us consider the TAU Request arrival diagram in Figure 30. As shown in the figure, *k* timeslots of *Δ* seconds of duration each have been distributed along the horizontal axis. Each of these timeslots may contain a *NAS Tracking Area Update Request* message (TAU Request), represented by a grey box in the figure.

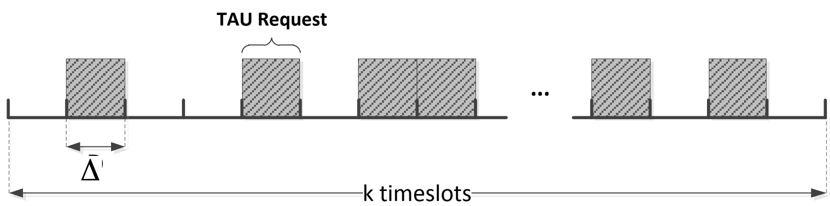


Figure 30.- TAU Request arrival diagram

The average number of TAU arrivals in k timeslots for a given UE ($\bar{N}_{TAU,k}$) can be calculated as follows:

$$\bar{N}_{TAU,k} = \sum_{i=1}^{k} p\{TAU, N\} = k \cdot p\{TAU, N\}$$

If the timeslot duration is Δ seconds, the average TAU arrival rate in a tracking area formed by N cell rings is:

$$\bar{\lambda}_{TAU,N} = \lim_{k\to\infty}\left(\frac{\bar{N}_{TAU,k}}{k \cdot \Delta}\right) = \frac{p\{TAU, N\}}{\Delta} \left[{}^{arrivals}/_{s}\right]$$

6.1.2 Mechanism Description

In order to minimise the impact on 3GPP Technical Specifications, the self-organising TAL mechanism builds upon the standard 3GPP TAU procedure defined in [5]. In particular, the P-MME entity monitors the arrival rate of *NAS Tracking Area Update Request* messages from the UE in order to estimate its mobility state (e.g., static, slow, medium, fast). Then, the P-MME updates the UE-specific Tracking Area List by increasing, maintaining, or reducing the number of cell rings in the TAL according to the estimated mobility state and the current paging arrival rate (note that the paging arrival rate can be easily computed locally at the P-MME). Finally, the P-MME sends the new TAL to the UE in the *NAS Tracking Area Update Accept* message.

The proposed self-organising mechanism combines the use of static and dynamic tracking area list mechanisms depending on the specific UE mobility state. Thus, TAL management remains static as long as location-related signalling is kept below a certain threshold. Once this *activation point* has been reached, the P-MME enables dynamic TAL management in order to reduce the overall location signalling traffic in the network. To further understand this behaviour, the concept of *stages* must be introduced. In stage 1, UEs are registered to Tracking Area Lists formed by a single cell ring. In stage 2, UEs are registered to TALs formed by two cell rings. In stage 3, TALs are formed by 3 cell rings, and so on.

In order to design an analytical model to study the proposed self-organising TAL mechanism, some assumptions must be made. First, we assume that the NoS is a hexagonal cell structure formed by concentric rings, as described in Figure 28. Secondly, we assume that all tracking areas in the NoS are formed by a single small cell. This enables the P-MME to treat small cells as TAs when managing Tracking Area Lists. Finally, we assume that UEs generate *NAS Tracking Area Update Request* messages according to a Poisson process with arrival rate $\bar{\lambda}_{TAU,i}$, where i denotes the number of cell rings in the i-th stage of the self-organising TAL mechanism.

We introduce two mobility management timers in the P-MME (T1, T2) to determine the mobility state of each UE. Timers are a common mechanism in 3GPP systems. Both UEs and EPC entities use timers to trigger signalling procedures, manage transitions between RRC states, monitor UE activity, release voice and data connections, control authentication protocols, etc. Some examples of 3GPP mobility management timers comprise T3412 (used to trigger periodic TAU procedures), T3311 (used to restart the attach procedure with the core network) or T303 (used to clear a voice call) [96]. In our proposal, a single (T1, T2) pair is used for each UE.

In each stage of the self-organising TAL mechanism, T1 and T2 are initialised to different values, namely *T1(i)* and *T2(i)*, where *i* denotes the corresponding stage. *T1(i)* controls transitions from the current to the next stage, while *T2(i)* controls transitions from the current to the previous stage. All possible transition events (NEXT_STAGE, CURRENT_STAGE, PREVIOUS_STAGE) are illustrated in Figure 31. The stage transition algorithm is described below:

- If the P-MME receives a *NAS Tracking Area Update Request* message before *T1(i)* expires, the self-organising TAL mechanism transits to the next stage. This corresponds to a mobility scenario where a UE is moving too fast for the corresponding TAL size in the current i-th stage.

- If the P-MME receives a *NAS Tracking Area Update Request* message after *T1(i)* expires and before *T2(i)* expires, the self-organising mechanism remains in the current stage. This corresponds to a scenario where a UE moves at an appropriate speed for its TAL size in the current i-th stage.

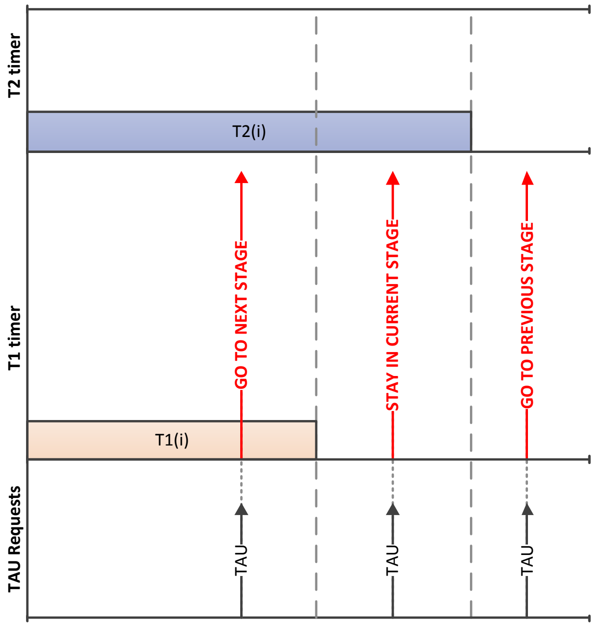


Figure 31.- Transition events in the self-organising TAL mechanism

- If the P-MME receives a *NAS Tracking Area Update Request* message after *T2(i)* expires, the self-organising mechanism transits to the previous stage. This corresponds to a scenario where a UE is moving too slowly for its corresponding TAL size in the current *i*-th stage.

After executing the stage transition algorithm, the P-MME updates the corresponding Tracking Area List locally and sends it to the UE encapsulated in the *NAS Tracking Area Update Accept* message, as described in [5].

The self-organising TAL mechanism can be mathematically described through a discrete-time Markov chain. States in the DTMC identify stages in the proposed mechanism. Analogously, transitions between states are associated with individual UE mobility events, such as staying in the current stage ($p\{stay, i\}$) or moving (forward or backwards) to an adjacent stage ($p\{next, i\}$ and $p\{previous, i\}$, respectively). Figure 32 shows the DTMC of a self-organising TAL mechanism with N stages.

In order to characterise the DTMC, we need to compute $p\{next, i\}$, $p\{previous, i\}$, $p\{stay, i\}$ $\forall i$ s.t. $1 \leq i \leq N$ (i.e., for all stages in the system). Since in each stage *NAS Tracking Area Update Request* arrivals follow a Poisson process with arrival rate $\bar{\lambda}_{TAU,i}$, the corresponding transition probabilities can be calculated as follows:

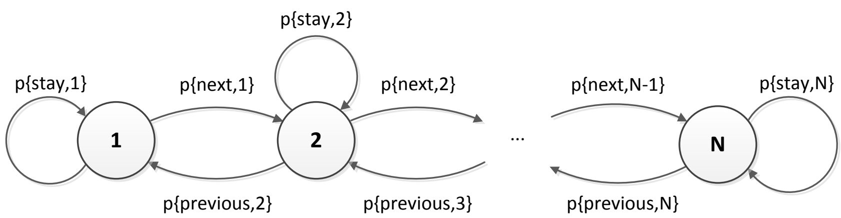


Figure 32.- DTMC of a self-organising TAL mechanism with N stages

$$p\{next, i\} = 1 - p\{no\ TAU\ arrivals\ during\ T1(i)\} = 1 - e^{-\lambda_{TAU,i} \cdot T1(i)}$$

$$p\{previous, i\} = p\{no\ TAU\ arrivals\ during\ T2(i)\} = e^{-\lambda_{TAU,i} \cdot T2(i)}$$

$$p\{stay, i\} = 1 - p\{next, i\} - p\{previous, i\}$$

Once all transition probabilities have been determined, we can calculate the one-step transition probability matrix of the DTMC, $\boldsymbol{P_{self}}$. As previously shown, the steady-state probability vector ($\boldsymbol{\pi_{self}}$) can be calculated as follows:

$$\boldsymbol{\pi_{self}} = [\pi_{self}(1) \ \dots \ \pi_{self}(N)] = \boldsymbol{e} \cdot (\boldsymbol{P_{self}} + \boldsymbol{E} - \boldsymbol{I})^{-1}$$

In conclusion, the self-organising Tracking Area List mechanism can be fully characterised by $\boldsymbol{\pi_{self}}$, as this vector contains the steady-state probabilities of finding the system in each one of the N stages.

6.1.3 Performance Evaluation

In this section we evaluate the performance of the self-organising TAL mechanism against that of a conventional (i.e., static) TAL mechanism, such as those traditionally deployed in current cellular networks. For both schemes, we define the following per-UE signalling cost function:

$$C_{tot} = p\{paging\} \cdot \bar{N}_{cells} \cdot c_p + p\{TAU\} \cdot c_{tau}$$

This function captures the signalling cost of the two fundamental location management procedures, i.e., paging and Tracking Area Update. Thus, $p\{paging\}$ is the probability of receiving a paging request in a given timeslot, $\bar{N}_{cells}$ is the average number of cells in the Tracking Area List (either static or dynamic) at the corresponding paging instant, c_p is the signalling cost of a single paging operation, $p\{TAU\}$ is the probability of receiving a *NAS Tracking Area Update Request* message in a given timeslot, and c_{tau} is the signalling cost of a single TAU operation. In general, a Tracking Area Update operation generates significantly more location-related signalling than a paging operation. In the performance evaluation of the self-organising TAL mechanism we have assumed a signalling ratio $\alpha = \frac{c_{tau}}{c_p} = 10$, which is a common value in the literature [97], [98], [99]. By normalising C_{tot} to c_p we obtain the following expression for the normalised signalling cost function:

$$C_{norm} = \frac{C_{tot}}{c_p} = p\{paging\} \cdot \bar{N}_{cells} + \alpha \cdot p\{TAU\}$$

In order to evaluate the performance of the proposed mechanism, we evaluate C_{norm} as a function of the UE speed for both the static and self-organising mechanisms. As described in Section 6.1.1, speed variations can be modelled by modifying the value of p in the Markov-based mobility model. Thus, low speeds correspond to values of p closer to 1, whilst high speeds correspond to values of p closer to 0.

The performance of the self-organising TAL mechanism depends on the values of the T1 and T2 timers in each stage of the system. In order to find the values that minimise the overall location-related signalling, we have used a sequential quadratic programming solver (SQP). The output of the SQP is a pair of vectors $(\mathbf{T1}_{opt}, \mathbf{T2}_{opt})$, where $\mathbf{T1}_{opt}$ and $\mathbf{T2}_{opt}$ contain the values of T1 and T2 that minimise C_{norm} in each stage of the self-organising mechanism. Thus:

$$\mathbf{T1}_{opt} = \left[\mathrm{T1}_{opt}(1)\ \ \mathrm{T1}_{opt}(2)\ \ \dots\ \ \mathrm{T1}_{opt}(N)\right]$$

$$\mathbf{T2}_{opt} = \left[\mathrm{T2}_{opt}(1)\ \ \mathrm{T2}_{opt}(2)\ \ \dots\ \ \mathrm{T2}_{opt}(N)\right]$$

Table 2 summarises the parameter values for the performance evaluation of the self-organising TAL mechanism. Numerical values for the paging arrival rate (λ_p) and signalling ratio (α) have been taken from the literature [47], [97], [98], [99]. The TAL size in the static mechanism (N_{stat}) has been set to 2 cell rings. Analogously, the maximum number of rings in a dynamic tracking area list (N_{dyn}) has been set to 4. The maximum small cell transmission radius (r) is 200m [20]. In our performance evaluation, UEs move throughout the NoS at speeds ranging from 0 to 50 km/h, thus covering both pedestrian and vehicular users. Finally, the timeslot duration can be calculated as follows:

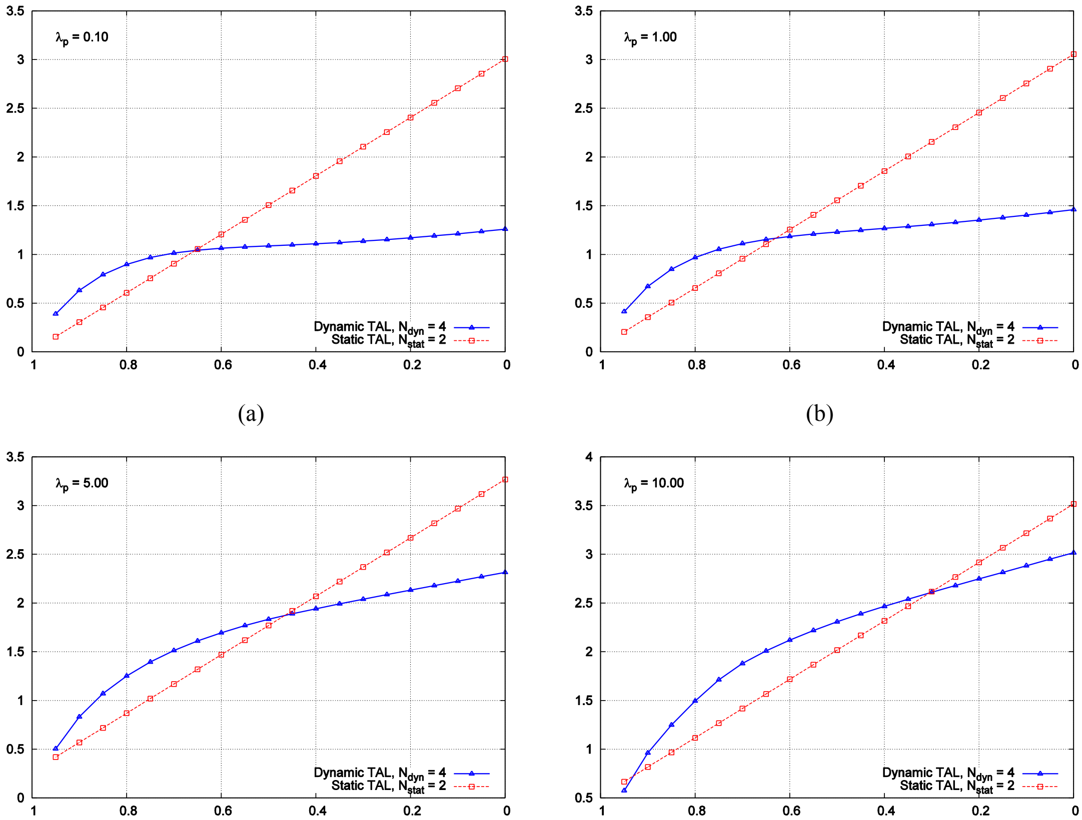
$$\Delta = \frac{2r}{v_{max}}$$

Table 2.- Performance evaluation parameters

Parameter	Description	Values
λ_p	Paging arrival rate [$pagings/h$]	[0.1 – 10]
α	Signalling ratio	10
N_{stat}	Number of rings in a static TAL	2
N_{dyn}	Maximum number of rings in a dynamic TAL	4
v	UE speed [km/h]	0 - 50
r	Small cell transmission radius [m]	200
Δ	Timeslot duration [s]	28.8

Figure 33 shows the impact of UE speed (represented in the X-axis by the parameter p introduced in Section 6.1.1) on the normalised signalling cost function (C_{norm}, Y-axis) for different values of paging arrival rate ($\lambda_p = \{0.1, 1, 5, 10\}\ pagings/h$). As shown in the figure, dynamic tracking area lists generate less location-related signalling than static TALs for medium- to high-speed UEs. This reduction is significantly higher when UEs are subject to moderate paging from the core network. Instead, for slow UEs, static tracking area lists generate less signalling traffic than dynamic TALs, as $p\{TAU\} \rightarrow 0$ and $\bar{N}_{cells,dyn} > \bar{N}_{cells,static}$. In general, since the cost of a single TAU operation is ten times that of a paging operation, the self-organising TAL mechanism aims at minimising $C_{norm}$by reducing the probability of TAU arrival for each specific UE.

The intersection of the solid and dotted curves in each figure determines the activation point of the self-organising TAL mechanism. Thus, at UE speeds where static TALs generate less location-related signalling than dynamic TALs, the proposed mechanism keeps the TAL size constant. Once the activation point has been reached, dynamic TAL management is enabled, hence reducing location-related signalling in the network. This strategy yields a significant reduction of location signalling traffic per UE, as shown in Table 3.



(a) (b)

(c) (d)

Figure 33.- Normalised signalling cost vs. p for different paging arrival rates

Table 3.- Location signalling reduction per UE

λ_p [pagings/h]	**Location Signalling Reduction**
0.1	39.45%
1	33.53%
5	13.21%
10	4.45%

Traditionally, MNOs use static TAL deployments in commercial networks according to proprietary network planning criteria. The number of cells per TA (and, subsequently, TAs per TAL) depends on various parameters, such as user density, traffic patterns, UE mobility, etc. Figure 34 compares the performance of dynamic TAL management against that of static TAL for different TAL sizes in a specific network scenario (other deployments show a similar generic behaviour). The arrows in the figure identify the activation points of the self-organising TAL mechanism. Depending on the scenario and the network planning criteria, the use of dynamic TAL sizes may benefit UEs for different speeds. In any case, our mechanism is designed to improve the performance of the static layout by enabling dynamic TAL management if doing so reduces location-related signalling.

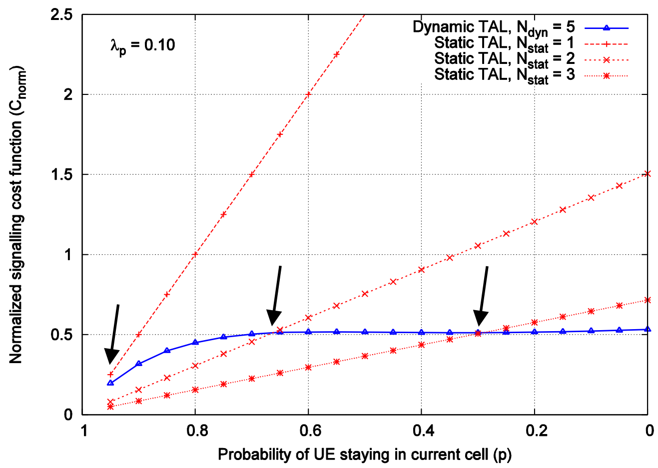


Figure 34.- Normalised signalling cost for different static TAL sizes

6.2 A Distributed Paging Mechanism over the X2 interface

The self-organising TAL mechanism described in the previous section addresses the limitations of static tracking area lists in the conventional 3GPP Tracking Area Update procedure. However, the issue of efficient paging in NoS remains unsolved. In fact, one of the biggest limitations of the self-organising TAL mechanism is the high volume of *S1-AP Paging* messages that need to traverse the wireless multi-hop backhaul of the NoS in order to reach a UE in high-mobility state. According to the mechanism description in Section 6.1.2, UEs in high-mobility states are registered to TALs that comprise a large number of small cells. Thus, whenever the network needs to page a fast UE, the P-MME needs to flood the wireless multi-hop backhaul with multiple *S1-AP Paging* messages in order to reach each and every small cell in the UE's current tracking area list. Flooding the network with location-related signalling increases the chances of network congestion in the wireless multi-hop backhaul.

In this section we propose a distributed paging mechanism over the X2 interface for all-wireless networks of small cells. Our scheme splits the standard 3GPP paging procedure into two stages in order to reduce the end-to-end paging delay and the total number of over-the-air (OTA) transmissions over the wireless multi-hop backhaul. In particular, the proposed mechanism sends a single unicast *S1-AP Paging* message from the P-MME to a selected small cell in the destination TA. Upon reception of this message, the receiving HeNB initiates a local paging procedure in the destination TA through the X2 interface until the destination UE has been paged. In order to avoid unnecessary flooding over the wireless multi-hop backhaul, the P-MME computes a bijective paging tree for the destination TA and encapsulates it in the S1- and X2-AP message payloads (this requires minor modifications to both protocols). Extensive simulation results show that the distributed paging mechanism reduces the total number of OTA transmissions in ~70% when compared to the standard 3GPP paging mechanism. In addition, our proposal also reduces the average end-to-end paging delay (i.e., the time it takes the EPC to page a UE) in ~50%, as well as the average number of accesses to the wireless multi-hop backhaul (MAC accesses) per byte (~80%).

6.2.1 Mechanism Description

The steps below describe the sequence of operations in the distributed paging mechanism, along with the corresponding S1-AP and X2-AP messages:

(a) Upon arrival of a voice or data connection request, the P-MME sends a single unicast *S1-AP Paging* message to a selected HeNB in the destination TA. The selected HeNB may be chosen based on various criteria (minimum distance to the P-MME, minimum congestion in the transport network, maximum number of one-hop neighbours in the vicinity of the selected HeNB, etc.). In the performance evaluation described in Section 6.2.3 we determined the selected HeNB based on minimum distance to the P-MME.

(b) In order to avoid unnecessary paging broadcasts in the destination TA, the P-MME computes a bijective paging tree prior to sending the unicast *S1-AP Paging* message to the selected HeNB. The paging tree contains a hierarchical paging sequence with the identities of the HeNBs that will be paged in the destination TA, along with the corresponding HeNBs in charge of paging them.

Essentially, the paging tree partitions the destination TA in disjoint branches such that each HeNB is paged only once.

(c) Once computed, the P-MME serialises the paging tree into a Prüfer sequence and encapsulates it in the *S1-AP Paging* message payload. Prüfer sequences are an efficient mechanism to encode and decode tree structures, as they minimise the number of entries in the serialised output and involve a low computational cost. A detailed description of Prüfer sequences can be found in [100].

(d) After receiving the unicast *S1-AP Paging* message from the P-MME, the selected HeNB sends a paging notification over the LTE-Uu interface. In parallel, it starts decoding the Prüfer sequence in the *S1-AP Paging* payload to reconstruct the paging tree. Once finished, the HeNB scans the paging tree, extracts the subtrees corresponding to each one of its one-hop neighbours, and computes their corresponding Prüfer sequences. Finally, the HeNB encapsulates these sequences into separate *X2-AP Paging* messages and sends them to its adjacent HeNBs via the X2 interface.

(e) Each neighbouring HeNB that receives an *X2-AP Paging* message over the X2 interface repeats step (d) until the incoming paging tree is empty.

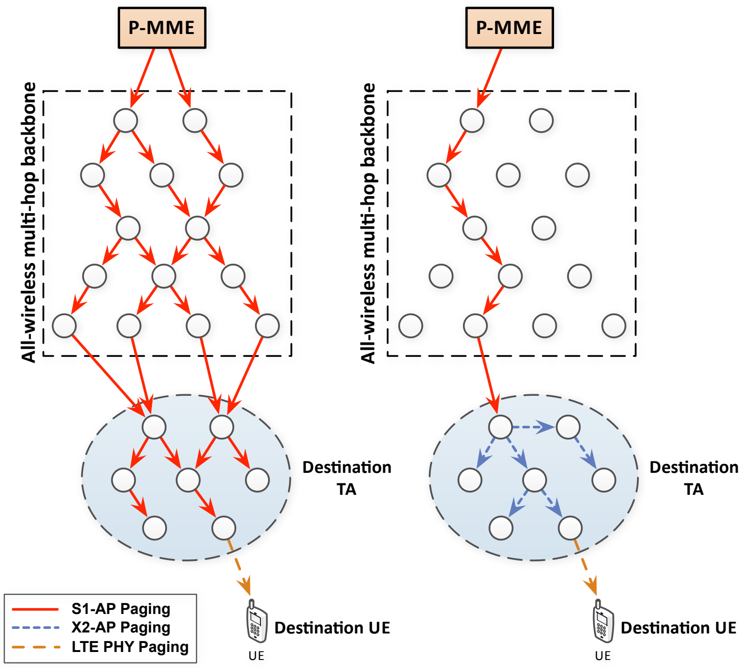


Figure 35.- Standard vs. distributed paging mechanism in a NoS

Figure 36 shows the sequence of protocol messages exchanged between EPS functional entities involved in the distributed paging mechanism. A detailed description of each step in the message sequence chart is provided below:

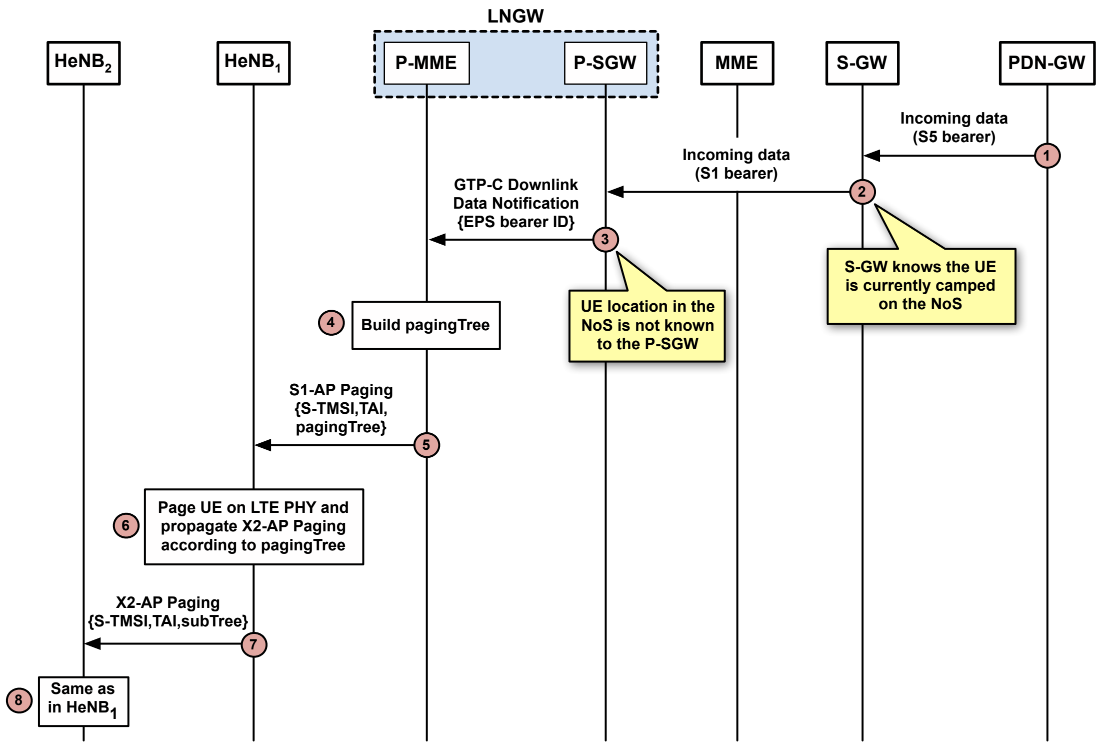


Figure 36.- Message sequence chart of the distributed paging mechanism

1. The PDN-GW in the Evolved Packet Core sends an incoming data packet to the S-GW over the S5 bearer. In practice, this indicates that a voice call or a data session needs to be established with a UE in the network of small cells.

2. The S-GW forwards the incoming packet to the P-SGW in the LNGW over a previously established S1 bearer. For the sake of simplicity, we have assumed that the S-GW knows that the UE is currently camped on the NoS. However, the exact location of the serving HeNB (and, thus, the UE) remains unknown.

3. The P-SGW sends a *GTP-C Downlink Data Notification* message to the P-MME. The purpose of this message is to notify the P-MME that incoming data is ready to be sent to the destination UE.

4. The P-MME obtains the destination tracking area from the *EPS bearer ID* field in the *GTP-C Downlink Data Notification* message and builds a paging tree with the selected HeNB as the root node. All HeNBs in the destination tracking area are included in the paging tree.

5. The P-MME encodes the paging tree into a Prüfer sequence and encapsulates it in the payload of a unicast *S1-AP Paging* message. This message is sent to the selected HeNB ($HeNB_1$), along with the UE SAE Temporary Mobile Subscriber Identity (S-TMSI) and the Tracking Area Identity (TAI).

6. The selected HeNB receives the *S1-AP Paging* message from the P-MME and sends a paging notification over the LTE-Uu interface, as defined in [101]. In parallel, it reconstructs the paging tree and extracts the subtrees rooted at each one of its children nodes (i.e., one-hop neighbouring HeNBs in the NoS). Later on, each subtree is serialised into a Prüfer sequence and encapsulated into a separate *X2-AP Paging* message (one for each child HeNB). This is done to reduce the size of the *X2-AP Paging* payload as paging messages are propagated throughout the wireless multi-hop backhaul.

7. The selected HeNB sends an *X2-AP Paging* message to each one-hop neighbouring HeNB, along with the paging subtree, the S-TMSI, and the TAI.

8. Upon reception of an *X2-AP Paging* message, a neighbouring HeNB repeats step 6 until an empty paging tree is received.

In order to encapsulate the Prüfer sequence that contains the paging tree (and subtrees), we define a new *Information Element* (IE) in the 3GPP S1-AP and X2-AP protocol specifications. Figure 37 shows the definition of the new *PagingTree* IE in Abstract Syntax Notation One (ASN.1). As far as protocol specification is concerned, we define a paging tree as a finite sequence of *E-UTRAN Cell Identifiers*, i.e., a standard 28-bit string used to identify (H)eNBs in the Evolved Packet System. In order to mitigate the impact on 3GPP Technical Specifications, the structure of the *PagingTree* IE is analogous to that of Tracking Area Lists, as defined in [101].

```
PagingTree ::=
  SEQUENCE (SIZE(1.. maxNoOfHeNBsInTree)) OF
  ProtocolIE-SingleContainer {{ EUTRANCellIdentifier }}
EUTRANCellIdentifier ::= BIT STRING (SIZE (28))
```

Figure 37.- ASN.1 definition of the PagingTree Information Element

6.2.2 Evaluation Metrics

This subsection defines the QoS metrics used in the performance evaluation of the proposed distributed paging mechanism. We define these metrics as follows:

- ***Number of OTA paging transmissions ($N_{tx,TOT}$):*** this metric captures the total number of S1-AP and X2-AP paging transmissions over the wireless multi-hop backhaul during a paging procedure. This includes MAC retransmissions caused by frame collisions and losses in the wireless channel. Thus:

$$N_{tx,TOT} = \begin{cases} N_{tx,S1AP}, & standard\ mechanism \\ (N_{tx,S1AP} + N_{tx,X2AP}), & distributed\ mechanism \end{cases}$$

- ***Paging payload (P_{paging}):*** this metric evaluates the total number of S1-AP and X2-AP payload bytes transmitted over the wireless multi-hop backhaul during the execution of a paging procedure. P_{paging} effectively captures the impact of the *PagingTree* IE on the *S1-AP* and *X2-AP Paging* messages payload.

$$P_{paging} = \begin{cases} P_{S1AP}, & standard\ mechanism \\ (P_{S1AP} + P_{X2AP}), & distributed\ mechanism \end{cases}$$

- ***Number of MAC accesses per paging payload byte:*** this metric captures the total number of accesses to the wireless multi-hop backhaul per (S1-AP + X2-AP) payload byte, including retransmissions. This has implications in energy efficiency and end-to-end paging delay, as each MAC access involves RF power consumption, inter-frame waiting periods, and potential delays due to contention mechanisms and retransmissions.

- ***Average end-to-end paging delay:*** this metric evaluates the average time it takes for a paging mechanism to complete a full paging operation. It is defined as the time elapsed between the arrival of the *GTP-C Downlink Data*

Notification message to the P-MME and the reception of the last *S1-AP/X2-AP Paging* message in the destination tracking area.

6.2.3 Performance Evaluation

This section discusses the performance evaluation of the distributed paging mechanism based on the metrics defined above. All results have been obtained after carrying out extensive simulations with the open source ns-3 network simulator [102].

6.2.3.1 Simulation Scenario

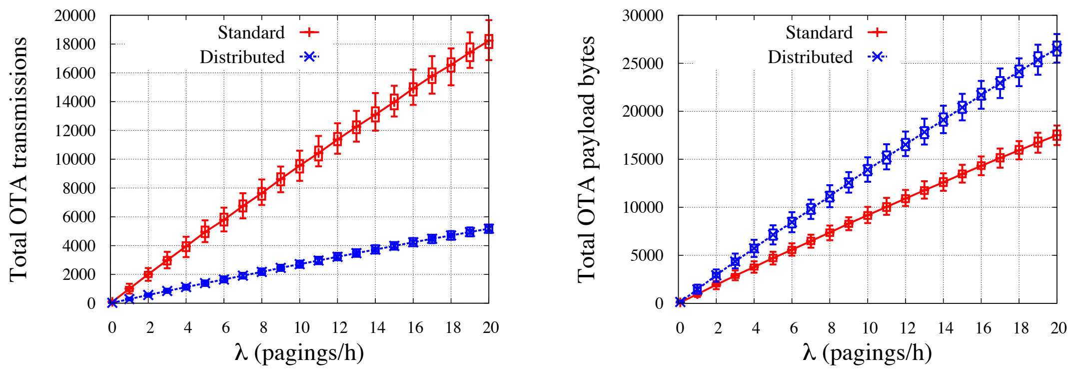
Initially, we consider a fully meshed, (6x6)-node wireless multi-hop backhaul with 35 HeNBs and a single P-MME located in the top-right corner of the network of small cells. Each HeNB can only communicate with its one-hop neighbours via the X2 interface. This logical interface is mapped to a physical 802.11g link between neighbouring HeNBs. We divide the NoS in 4 identical tracking areas, each one comprising a subset of 3x3 HeNBs. UEs move throughout the NoS following a Gauss-Markov mobility model [95] with speeds ranging from 5 to 50 km/h. *GTP-C Downlink Data Notification* messages arrive to the P-MME entity following a Poisson process with paging arrival rate $\lambda \in [0.1, 20]$ pagings/h. In order to average the simulation results, we ran 100 repetitions for each paging scheme (standard vs. distributed). The duration of each repetition was 100000 seconds. Table 4 summarises the simulation parameters for the performance evaluation of the distributed paging mechanism.

Table 4.-Simulation parameters for performance evaluation

Name	Description	Value
λ	Paging arrival rate [*pagings/h*]	[0.1 - 20]
N	Total number of nodes	35 HeNBs 1 P-MME
$N_{small\_cells,TA}$	Number of HeNBs per TA	9
v	UE speed [*km/h*]	5 - 50
N_r	Number of repetitions	100
T	Repetition duration [*s*]	100000

6.2.3.2 Simulation Results

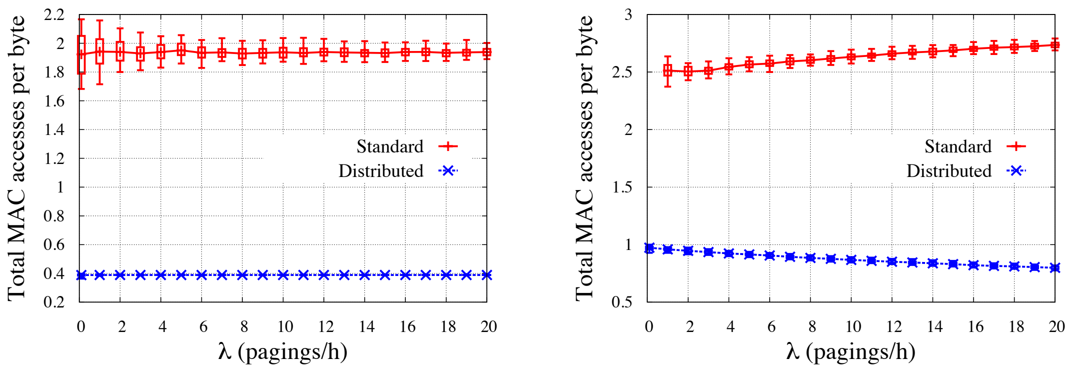
Figure 38(a) shows the total number of OTA paging transmissions (S1-AP and X2-AP) over the wireless multi-hop backhaul in the standard and distributed paging mechanisms. As shown in the figure, the number of OTA transmissions is consistently lower in the distributed mechanism for all values of λ. In addition, the variance of the number of transmissions is also lower in the distributed scheme, and remains bounded as λ increases. This is a consequence of the distributed paging algorithm. By sending a single *S1-AP Paging* message to the selected HeNB followed by one-hop *X2-AP Paging* messages in the destination tracking area, the number of OTA paging transmissions needed to reach all HeNBs is significantly reduced. In addition, this has a positive impact on network congestion, thus lowering channel contention, collisions, and frame retransmissions.



(a) (b)

Figure 38.- OTA paging transmissions (a) and OTA paging payload bytes (b)

Figure 38 (b) shows the impact of the *PagingTree* Information Element on the *S1-AP* and *X2-AP Paging* payloads. As shown in the figure, the total number of OTA payload bytes in the distributed mechanism is higher than in the standard mechanism. This is due to the encapsulation of the Prüfer sequences containing the paging tree (and subtrees) in the S1-AP and X2-AP payload, respectively. In both schemes, the variance in the number of OTA payload bytes is caused by frame retransmissions. However, the impact of these retransmissions is different depending on the selected paging mechanism. In the standard scheme, only *S1-AP Paging* messages are retransmitted. These messages always contain a single *E-UTRAN Cell Identifier* (28 bits) in order to determine the target HeNB in the destination tracking area. Instead, in the distributed paging mechanism, both *S1-AP* and *X2-AP Paging* messages may be retransmitted. The *S1-AP Paging* payload is fixed, since it always contains the entire paging tree. However, the *X2-AP Paging* payload is variable, as it depends on the size of the extracted subtree. This is the reason why the variance in the distributed mechanism is higher than in the standard mechanism.



(a) (b)

Figure 39.- MAC accesses/byte without (a) and with (b) background traffic

Figure 39 shows the ratio between the number of MAC accesses (including retransmissions) and the S1-AP plus X2-AP payload bytes needed to complete a paging operation. In order to evaluate this metric under different network conditions, we have defined two traffic scenarios. In the first setup, we have reduced background traffic in the wireless backhaul by pre-populating the Address Resolution Protocol (ARP) tables for all HeNBs in the network of small cells. We refer to this setup as the *background traffic-free scenario*. In the second setup, we have enabled ARP traffic between HeNBs to increase congestion, thus triggering channel contention, collisions and retransmissions. This setup is referred to as the *background traffic scenario*.

Figure 39(a) shows the number of accesses to the wireless multi-hop backhaul per payload byte in the background traffic-free scenario. As seen in the figure, our proposal outperforms the standard mechanism in terms of MAC accesses per byte for all paging arrival rates. To this extent, the lower number of over-the-air transmissions needed to complete a paging procedure in the distributed mechanism compensates the bigger payload due to the paging tree encapsulation. Analogously, Figure 39(b) shows the number of MAC accesses per payload byte in the background traffic scenario. As seen in the figure, the presence of ARP traffic in the wireless multi-hop backhaul has an impact on the performance of both paging mechanisms. In general, our proposal generates less ARP traffic than the standard mechanism due to the lower number of accesses to the wireless multi-hop backhaul needed to complete a full paging operation. This has a positive impact on the number of MAC accesses per payload byte, as less background ARP traffic means less collisions and frame retransmissions. On the other hand, the standard paging mechanism generates more ARP traffic due to the higher number of MAC accesses needed to complete a full paging operation. This increases congestion in the wireless multi-hop backhaul, thus resulting in a higher number of MAC accesses per payload byte. In both figures, the higher variance in the standard mechanism for $0 \leq \lambda \leq 6$ is caused by the sparse arrival of *GTP-C Downlink Data Notification* messages during simulation time (i.e., fewer samples). This effect is less noticeable as the paging arrival rate increases.

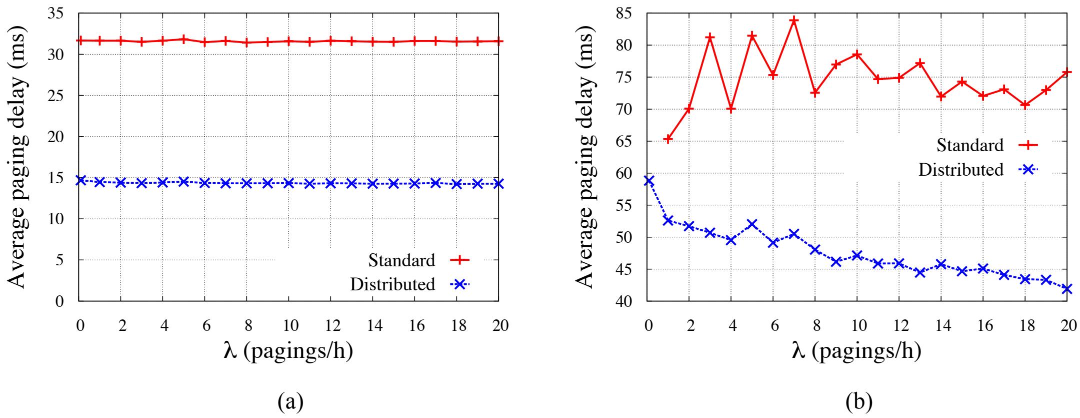


Figure 40.- Average paging delay without (a) and with (b) background traffic

Figure 40 shows the end-to-end paging delay in the standard and distributed paging mechanisms for both background traffic scenarios. For clarity purposes, only average values over all simulation repetitions have been depicted. As shown in Figure 40(a), the distributed paging mechanism reduces the average paging delay by 50% compared to the standard mechanism in the absence of background ARP traffic. This is due to the lower number of multi-hop paging transmissions needed to reach all HeNBs in the destination tracking area. In the standard paging mechanism, each *S1-AP Paging* message needs to traverse the multi-hop wireless backhaul before reaching the destination HeNB. This involves channel contention and transmission delays, as well as potential collisions and frame retransmissions, hence increasing the paging delay. Instead, in the distributed paging scheme, the P-MME entity reduces the number of transmissions (and, therefore, contention and collisions) in the wireless channel by sending a single *S1-AP Paging* message to a selected HeNB in the destination TA. All subsequent *X2-AP Paging* messages between neighbouring HeNBs in the destination TA are (by definition) one-hop transmissions. This reduces channel contention, as well as collisions and frame retransmissions, thus resulting in lower end-to-end paging delays.

As seen in Figure 40(b), the presence of background ARP traffic in the wireless multi-hop backhaul increases the end-to-end paging delay for both paging mechanisms. In general, increasing the network load leads to more contention in the wireless channel. Eventually, this leads to more collisions, frame retransmissions, and higher end-to-end paging delays. Furthermore, additional delays introduced by the ARP mechanism are a key contributing factor to the overall end-to-end paging delay. In general, MAC entries in ARP tables are more likely to expire for lower values of paging arrival rates. Consequently, HeNBs need to re-populate their ARP tables every time a *GTP-C Downlink Data Notification* message reaches the P-MME, thus resulting in higher end-to-end paging delays. As λ increases, voice/data connection requests reach the P-MME entity before MAC entries in the ARP tables expire, which prevents HeNBs from re-initiating ARP mechanisms in the network.

6.2.4 Scalability Evaluation

In this section we study the scalability of the distributed paging mechanism, i.e., the evaluation of the performance metrics defined in Section 6.2.2 as a function of the tracking area size ($N_{small\_cells,TA}$).

6.2.4.1 Simulation Scenario

In order to evaluate the scalability of the distributed paging mechanism, we define a new simulation scenario. In particular, we consider a fully meshed (24x24)-node wireless multi-hop network with 575 HeNBs and one P-MME located in the centre of the network of small cells. Each HeNB can only communicate with its one-hop neighbours via an X2 interface mapped to a physical 802.11g link. As far as the tracking area size is concerned ($N_{small\_cells,TA}$), we consider six different values, namely 1, 4, 9, 16, 36, and 64 HeNBs per TA. UEs move throughout the NoS following a Gauss-Markov mobility model with speeds ranging from 5 to 50 km/h. *GTP-C Downlink Data Notification* messages reach the P-MME following a Poisson process with paging arrival rate λ = 10 pagings/h. In order to average the simulation results, we ran 50 repetitions for each paging scheme (standard vs. distributed). The duration of each repetition was 10000 seconds. These simulation parameters are summarised in Table 5.

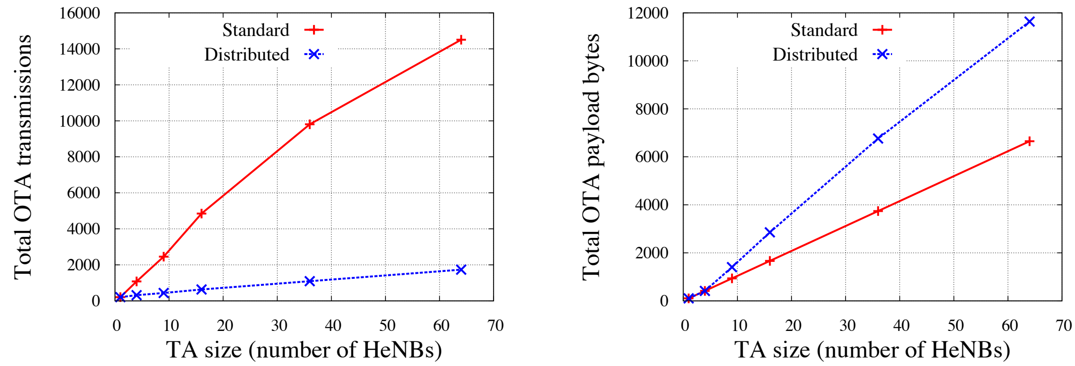
Table 5.-Simulation parameters for scalability evaluation

Name	Description	Value
λ	Paging arrival rate [*pagings/h*]	10
N	Total number of nodes	575 HeNBs, 1 P-MME
$N_{small\_cells,TA}$	Number of HeNBs per TA	1, 4, 9, 16, 36, 64
v	UE speed [*km/h*]	5 - 50
N_r	Number of repetitions	50
T	Repetition duration [*s*]	10000

6.2.4.2 Results

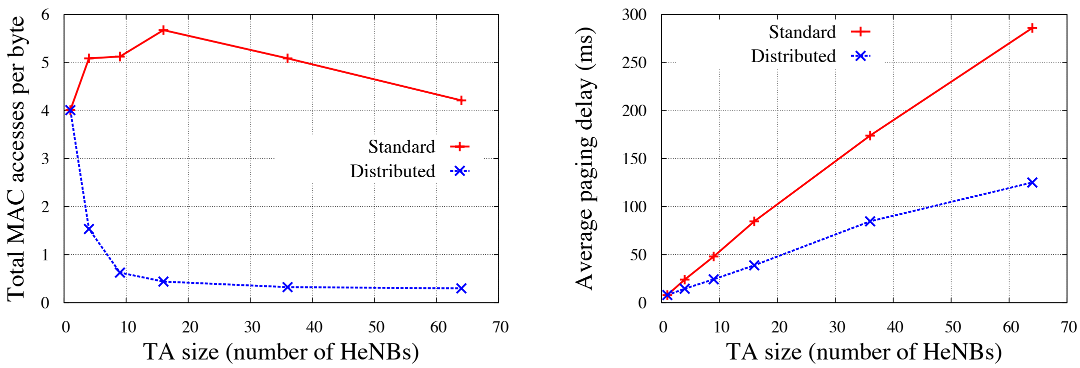
Figure 41(a) shows the total number of OTA transmissions needed to complete a full paging procedure for different TA sizes. As shown in the figure, the distributed mechanism scales better than the standard mechanism for large deployments of all-wireless NoS. As the total number of HeNBs increases, each *S1-AP Paging* message needs to traverse a higher number of hops in the NoS in order to reach its destination HeNB. In the standard mechanism, the total number of OTA transmissions needed to complete a full paging procedure scales by the number of hops of a single *S1-AP Paging* message times the number of HeNBs in the destination TA. However, in the distributed mechanism, the total number of OTA transmissions needed to complete a paging procedure scales by the number of HeNBs in the destination TA only.

Figure 41(b) shows the total number of OTA payload bytes in *S1-AP* and *X2-AP Paging* messages for different TA sizes. As expected, the number of payload bytes in the distributed paging mechanism is higher than in the standard one due to the encapsulation of the paging tree. Thus, in the standard paging scheme, the total number of OTA payload bytes scales by the number of *S1-AP Paging* messages times the size of a single *E-UTRAN Cell Identifier*. However, in the distributed paging scheme, the total number of OTA payload bytes needed to complete a full paging operation scales by the size of the paging tree/subtrees and the total number of *S1-AP* and *X2-AP Paging* messages.



(a) (b)

Figure 41.- OTA transmissions (a) and payload bytes (b) for different TA sizes



(a) (b)

Figure 42.- MAC accesses/paging payload byte (a) and average paging delay (b)

Figure 42(a) shows the total number of accesses to the wireless multi-hop backhaul (MAC accesses) per payload byte for different tracking area sizes. As shown in the figure, the distributed paging mechanism is more efficient than the standard paging mechanism. This is achieved by packing a higher number of bytes (i.e., the paging tree) in fewer OTA transmissions. The positive impact of the paging tree in the number of total MAC accesses per byte becomes more noticeable as the number of HeNBs in the destination TA increases. Essentially, this means that for large deployments of NoS is more efficient to pack multiple HeNB identifiers in a single paging tree than sending separate *S1-AP Paging* messages to each HeNB in the destination tracking area.

Figure 42(b) shows the average paging delay in each paging mechanism for different tracking area sizes. As previously discussed, channel contention, collisions, and frame retransmissions are the key contributing factors to end-to-end paging delay. Since the distributed paging mechanism reduces congestion in the wireless multi-hop backhaul, a lower number of *S1-AP* and *X2-AP Paging* messages experience channel contention and, eventually, collisions and frame retransmissions. Consequently, the positive impact of lower network congestion on end-to-end paging delay becomes more noticeable as the size of the destination tracking area increases.

6.3 Research Contributions

In the course of the design and evaluation of location management procedures for large-scale networks of small cells, the following contributions to the literature have been made:

- J. Ferragut, J. Mangues-Bafalluy, "A Distributed Paging Mechanism over the X2 Interface for All-Wireless Networks of Small Cells", in Proc. 7th IFIP Wireless and Mobile Networking Conference (WMNC), 20-22 May, 2014, Vilamoura (Portugal). Acceptance rate (2014): 32.9%

- J. Mangues, J. Ferragut, M. Requena, "Distributed Location Management for Generalized HetNets. Case Study of All-Wireless Networks of Femtocells", Heterogeneous Cellular Networks, edited by R.Q. Hu and Y. Qian, published by John Wiley & Sons Ltd, Oxford, UK, April 2013.

- J. Ferragut, J. Mangues-Bafalluy, J. Núñez, F. Zdarsky, "Traffic and Mobility Management in Networks of Femtocells", ACM/Springer Mobile Networks and Applications Journal, Special Issue on Cooperative and Networked Femtocells, Vol. 2012, pp. 662-673, August 2012. Impact factor (2012): 1.109

- J. Ferragut, J. Mangues-Bafalluy, "A Self-Organized Tracking Area List Mechanism for Large-Scale Networks of Femtocells", in Proc. IEEE International Conference on Communications (ICC), 10-15 June, 2012, Ottawa (Canada). Acceptance rate (2012): 37%

- INFSO-ICT-248523 BeFEMTO – Deliverable D2.2, v1.0, "The BeFEMTO System Architecture" [Online]. Available: http://www.ict-befemto.eu

- INFSO-ICT-248523 BeFEMTO – Deliverable D2.3, v1.0, "The BeFEMTO System Concept and its Performance" [Online]. Available: http://www.ict-befemto.eu
- INFSO-ICT-248523 BeFEMTO – Deliverable D5.1, v1.0, "Femtocell access control, networking, mobility, and management concepts" [Online]. Available: http://www.ict-befemto.eu
- INFSO-ICT-248523 BeFEMTO – Deliverable D5.2, v1.0, "Femtocell access control, networking, mobility and management mechanisms (final)" [Online]. Available: http://www.ict-befemto.eu
- INFSO-ICT-248523 BeFEMTO – Deliverable D5.3, v1.0, "Evaluation report of femtocells networking, mobility and management solutions" [Online]. Available: http://www.ict-befemto.eu

Chapter 7

Traffic Management Mechanisms in Networks of Small Cells

The use of efficient unicast and multicast traffic management mechanisms is a critical issue in large-scale deployments of networks of small cells. In these mechanisms, efficiency is measured in terms of radio and network resource usage within the NoS and over the backhaul link towards the Evolved Packet Core. This is particularly important for all-wireless networks of small cells deployments, where connectivity amongst neighbouring HeNBs (and, in turn, with the LNGW) is provided via a wireless multi-hop backhaul. In these scenarios, avoiding routing bottlenecks (e.g., the P-SGW entity in the LNGW) by ensuring an even distribution of resource consumption throughout the entire local network is paramount.

In this chapter we present the research work carried out on the issue of unicast and multicast traffic management mechanisms for all-wireless networks of small cells. Section 7.1 describes the necessary architectural and procedural framework to support the establishment of direct unicast bearers (also known as *D-bearers*) between UEs camped on the same network of small cells, though not necessarily in the vicinity of each other. These mechanisms must support the exchange of user-plane data between LTE peers through the LTE-Uu interface in both static and handover scenarios with minimal involvement from EPC entities. In addition, direct bearers enable the deployment of distributed routing protocols in the underlying transport network of a NoS, such as those described in [89]. These mechanisms ensure a fair distribution of radio and network resource consumption throughout the entire network of small cells, thus reducing user-plane traffic on the backhaul link towards the EPC and avoiding potential bottlenecks.

As far as multicast traffic management is concerned, networks of small cells constitute a flexible and cost-effective solution for MNOs to support MBMS services in localised areas, such as convention centres, sports venues, transportation hubs, university campuses, shopping malls, financial districts, etc. However, standard 3GPP MBMS services suffer from the traditional limitations of wireless multicast/broadcast traffic distribution systems, namely the use of robust modulation and coding schemes (and, thus, low bit rates) in order to guarantee an acceptable QoE for cell-edge users experiencing the worst radio conditions. In order to overcome these limitations, we propose a RAT-agnostic feedback collection mechanism between MBMS UEs and their corresponding Multi-Cell/Multicast Coordination Entities (MCEs). Our proposal builds upon the concept of *feedback selection* and *rate adaptation* to dynamically adjust the MBMS modulation and coding scheme to the available radio and network resources whilst keeping signalling traffic over the LTE-Uu interface under control. An experimental evaluation of the lightweight feedback selection and rate adaptation mechanism, along with its implications in the MBMS network and protocol stack architecture, is described in detail in Section 7.2.

7.1 Unicast Traffic Management

In all-wireless networks of small cells, connectivity amongst HeNBs is provided via a wireless multi-hop transport network. Thus, in addition to the standard 3GPP LTE-Uu interface (which provides UEs in the NoS with access to the E-UTRAN), HeNBs in a NoS are also equipped with an additional wireless interface to support connectivity with neighbouring small cells. The underlying wireless technology in this additional interface is vendor-specific. Some options comprise conventional Wi-Fi (802.11abgn), carrier-grade Wi-Fi, microwave links, LTE self-backhauling, etc.

Building a large-scale wireless multi-hop backhaul amongst HeNBs enables the deployment of distributed routing protocols for either local (where the source and destination nodes are camped on the same network of small cells) or external traffic (where the destination node is located outside the NoS). The basis of these routing protocols is the avoidance of predetermined routes between the source and destination nodes. Instead, routing decisions are computed on a hop-by-hop basis according to the destination address and the state of the routing queues in each node. Distributed routing protocols are particularly suitable for large-scale deployments of all-wireless networks of small cells, as they require low control overhead (generally, a few HELLO messages between neighbouring HeNBs) and no centralised routing entity.

A particular example of a distributed and stateless routing protocol for all-wireless networks of small cells was developed in the context of the European Project BeFEMTO in conjunction with the traffic and mobility management mechanisms presented in this Ph.D. Thesis. For a detailed description of the design, development, simulation, and experimental evaluation of this routing protocol instantiation (from the transport network layer's perspective) the reader is referred to [89]. Instead, the following sections discuss the necessary 3GPP control-plane network procedures to support the establishment of direct IP flows in a NoS under the assumption that the distributed and stateless routing protocol described in [89] has been deployed in the underlying transport network. In particular, these procedures comprise both static and dynamic scenarios (i.e., where both the source and destination UEs are static or, alternatively, where one of the two UEs (or both) is performing a handover between neighbouring small cells).

7.1.1 D-Bearers

As discussed in Section 4.3, direct bearers (D-bearers) enable the exchange of unicast user-plane data between UEs camped on the same network of small cells (though not necessarily in the proximity of each other) without the need of centralised routing via the LNGW. D-bearers aim at fully exploiting the possibilities of a wireless multi-hop backhaul combined with a distributed and stateless routing protocol in the transport network layer. It is important not to confuse D-bearers with LIPA services. The former provide a QoS-enforcing mechanism for direct multi-hop user-plane data exchange between UEs camped on the same NoS without the need for a centralised routing element (i.e., an alternative approach to the standard 3GPP EPS bearer described in Section 2.4.1.1). On the other hand, LIPA allows IP-capable UEs camped on a standalone small cell to communicate with other IP-capable entities connected to the same residential/enterprise network, such as a file server, a printer, a local host, etc. Essentially, LIPA assumes an asymmetric communication scenario, where a UE connects to a HeNB via the LTE-Uu interface in order to reach a remote host located in a local IP-based network. Instead, D-bearers are (by definition) fully symmetrical, as they allow the establishment of a direct user-plane bearer via the LTE-Uu interface between LTE peers camped on the same network of small cells.

In a centralised routing scenario (i.e., without D-bearer support), local and external user-plane traffic in a network of small cells must traverse the P-SGW entity in the LNGW. This leads to inefficient routing paths between the source and destination nodes and increases the chances of congestion in the network of small cells, particularly in the proximity of the LNGW network element. Instead, D-bearers enable direct exchange of user-plane data between UEs in the NoS without the need of centralised routing points, as shown in Figure 43.

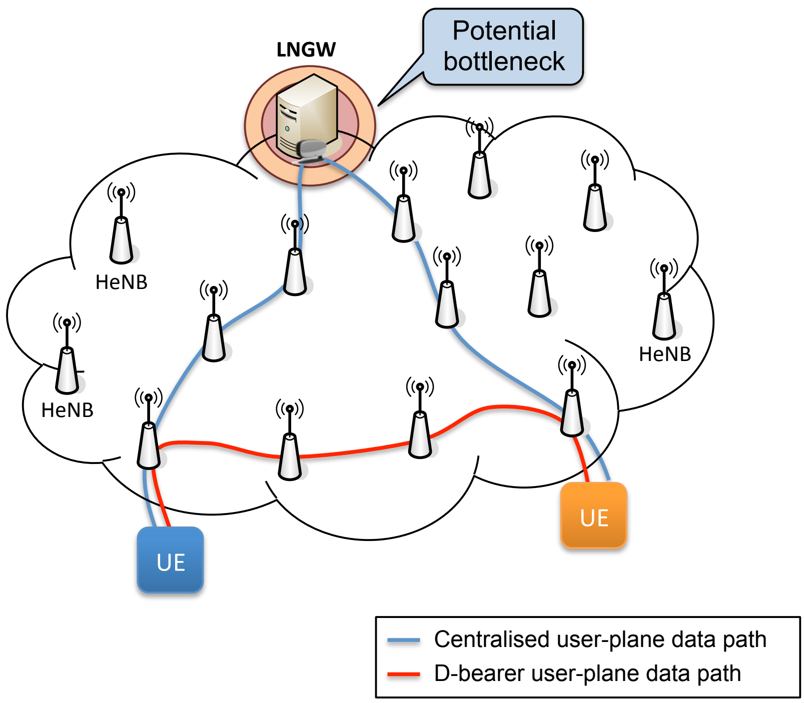


Figure 43.- Centralised vs. D-bearer data paths in a NoS scenario

In order to provide support for D-bearers, the EPS Bearer Service architecture described in Section 2.4.1.1 must be modified accordingly. In particular, this entails (a) designing a new network procedure aimed at establishing direct user-plane bearers between LTE peers camped on the same network of small cells and (b) supporting traffic scenarios where the source and/or destination UE(s) perform a handover whilst being engaged in user-plane data exchange over a D-bearer. In terms of QoS, direct bearers must ensure that the corresponding end-to-end QoS requirements are properly enforced between the radio network layer ingress and egress points (i.e., the source and destination UEs). In order to do so, EPS functional entities (UEs, HeNBs, and the P-MME) must be able to exchange the necessary control-plane messages to allocate the corresponding user-plane resources in the transport network layer. In practice, this involves translating end-to-end QoS requirements into the corresponding distributed routing protocol settings (e.g., priority bits, expedite processing policies, queue selection parameters, etc.). Section 7.1.3 describes the network procedure for the establishment of D-bearers in networks of small cells. On the other hand, handover support for D-bearers is discussed in Section 7.1.4.

7.1.2 The D-Bearer Service Architecture

Figure 44 depicts the architecture of a D-bearer established between two UEs camped on the same network of small cells. As shown in the figure, the end-to-end D-bearer (analogous to an EPS bearer in a conventional UE-to-PDN-GW scenario) comprises three *constituent bearers*, namely (a) a radio bearer between the source UE and its serving HeNB, (b) an inter-HeNB bearer between the source and the destination HeNBs (referred to as the *H-bearer*), and (c) a radio bearer between the destination UE and its serving HeNB. These bearers are described as follows:

- **Source radio bearer:** a standard user-plane radio bearer must be established between the source UE and its serving HeNB to support the exchange of data packets over the LTE-Uu interface according to a certain QoS level (i.e., RLC and MAC layer configuration). The source radio bearer is established by a conventional UE-initiated bearer establishment procedure (Section 2.4.1.4).

- **Inter-HeNB bearer (H-bearer):** the source and destination HeNBs involved in direct user-plane communication must establish a H-bearer between them in order to support the transport user-plane packets according to a certain QoS requirements. In turn, these requirements are translated into the corresponding configuration (e.g., scheduling priorities, queue selection, etc.) in the underlying transport network layer. From the point of view of the source HeNB, the H-bearer is established by triggering a conventional S1 bearer-establishment procedure towards the P-SGW in the LNGW. However, the P-MME detects that a direct user-plane bearer is being established and, in turn, takes the necessary steps to terminate the H-bearer in the destination HeNB instead of in the P-SGW. This procedure is described in detail in Section 7.1.3.

- **Destination radio bearer:** the destination UE and its serving HeNB must also establish a conventional user-plane radio bearer in order to transport IP packets over the E-UTRAN according to a predefined QoS requirements over the LTE-Uu interface. As described in the next section, the destination radio bearer is established as a result of a conventional paging procedure initiated by the P-MME. As seen in Figure 44, the destination radio bearer must be connected to the H-bearer at $HeNB_{dst}$ in order to enable the routing of IP packets between the source and destination UEs.

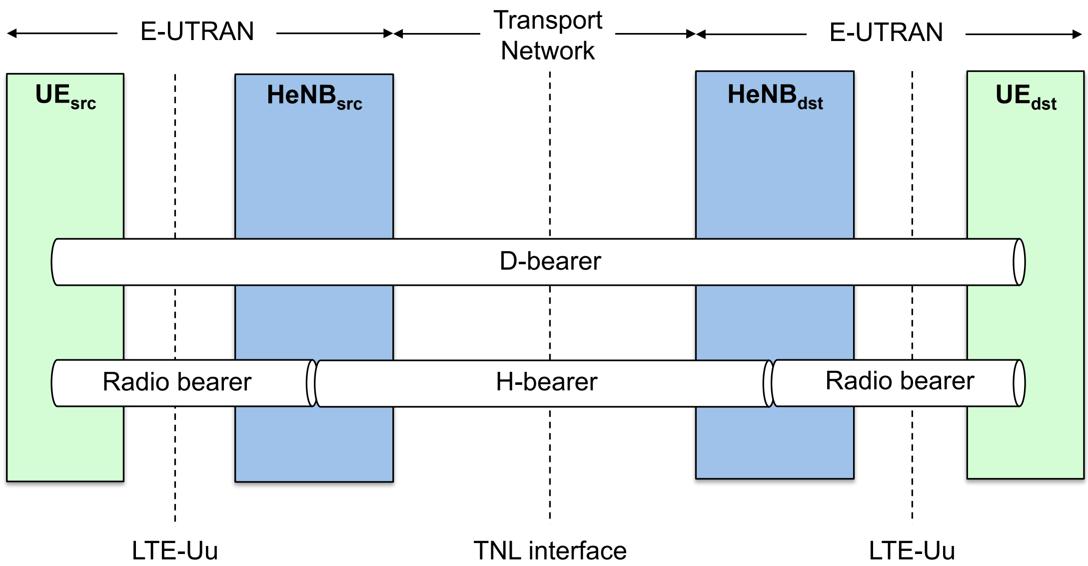


Figure 44.- The D-bearer architecture

As explained in Section 2.4.1.1, a *bearer* is a traffic management abstraction aimed at enforcing a certain set of QoS requirements over an IP flow between peer EPS functional entities. The bearer's ingress and egress points (e.g., the UE and the HeNB in the radio bearer, the source and destination HeNBs in the H-bearer, etc.) are the EPS entities responsible for enforcing the requested QoS by applying the appropriate settings at the corresponding layers in the protocol stack (e.g., RLC, MAC, PDCP, IP, etc.). Thus, a D-bearer is essentially an EPS mechanism aimed at enforcing a given set of QoS requirements over an IP flow established between a pair of LTE peers camped on the same network of small cells. In practice, the requested QoS is mapped onto specific RLC, MAC, and PDCP settings over the LTE-Uu interface (both at the source and destination radio links), as well as onto the corresponding configuration parameters in the underlying transport network.

7.1.3 The D-Bearer Establishment Procedure

Figure 45 shows the message sequence chart for the establishment of a D-bearer between two UEs in a network of small cells. As shown in the figure, UE_{src} initiates the direct bearer establishment by triggering a standard UE-initiated EPS bearer establishment procedure against the P-MME. As described in Section 2.4.1.4, this procedure establishes an EPS bearer between the UE and the PDN-GW in the Evolved Packet Core (e.g., to communicate with an Internet server). However, when the P-MME detects that UE_{src} is requesting a direct bearer with UE_{dst} (upon reception of the *S1-AP Initial UE message*) a standard paging procedure towards UE_{dst} is initiated instead, as depicted in the grey box in Figure 45. The cornerstone of the D-bearer establishment procedure is that UE_{src} behaves in the same way as a UE that is establishing a standard EPS bearer with the PDN-GW, whilst –on the other hand– UE_{dst} acts as a UE that is being paged by a P-MME. Thus, the P-MME acts as a proxy signalling entity between both UEs (and their corresponding HeNBs) in order to coordinate the sequence of control-plane messages and to allocate the necessary radio and network resources at each EPS functional entity. A detailed description of each step in the D-bearer establishment procedure is provided below:

- **Step 1:** UE_{src} establishes a signalling radio bearer with $HeNB_{src}$ in order to request a bearer resource allocation to the P-MME.

- **Steps 2 – 3:** UE_{src} sends a *NAS Bearer Resource Allocation Request* message to $HeNB_{src}$. In turn, this message is transparently forwarded to the P-MME in an *S1-AP Initial UE* message. The presence of UE_{dst}'s IP address in the TFT Information Element allows the P-MME to detect that UE_{src} is requesting the establishment of a direct bearer with UE_{dst}.

- **Step 4:** at this stage, the P-MME is aware that UE_{src} is attempting to establish a direct bearer with UE_{dst}. Thus, the P-MME initiates a standard paging procedure towards UE_{dst} by sending an *S1-AP Paging* message to all HeNBs in UE_{dst}'s last registered tracking area.

- **Steps 5 – 6:** $HeNB_{dst}$ sends an *RRC Paging* message to UE_{dst}. UE_{dst} responds by establishing an RRC connection with $HeNB_{dst}$ prior to executing a standard network-initiated bearer establishment procedure with the P-MME.

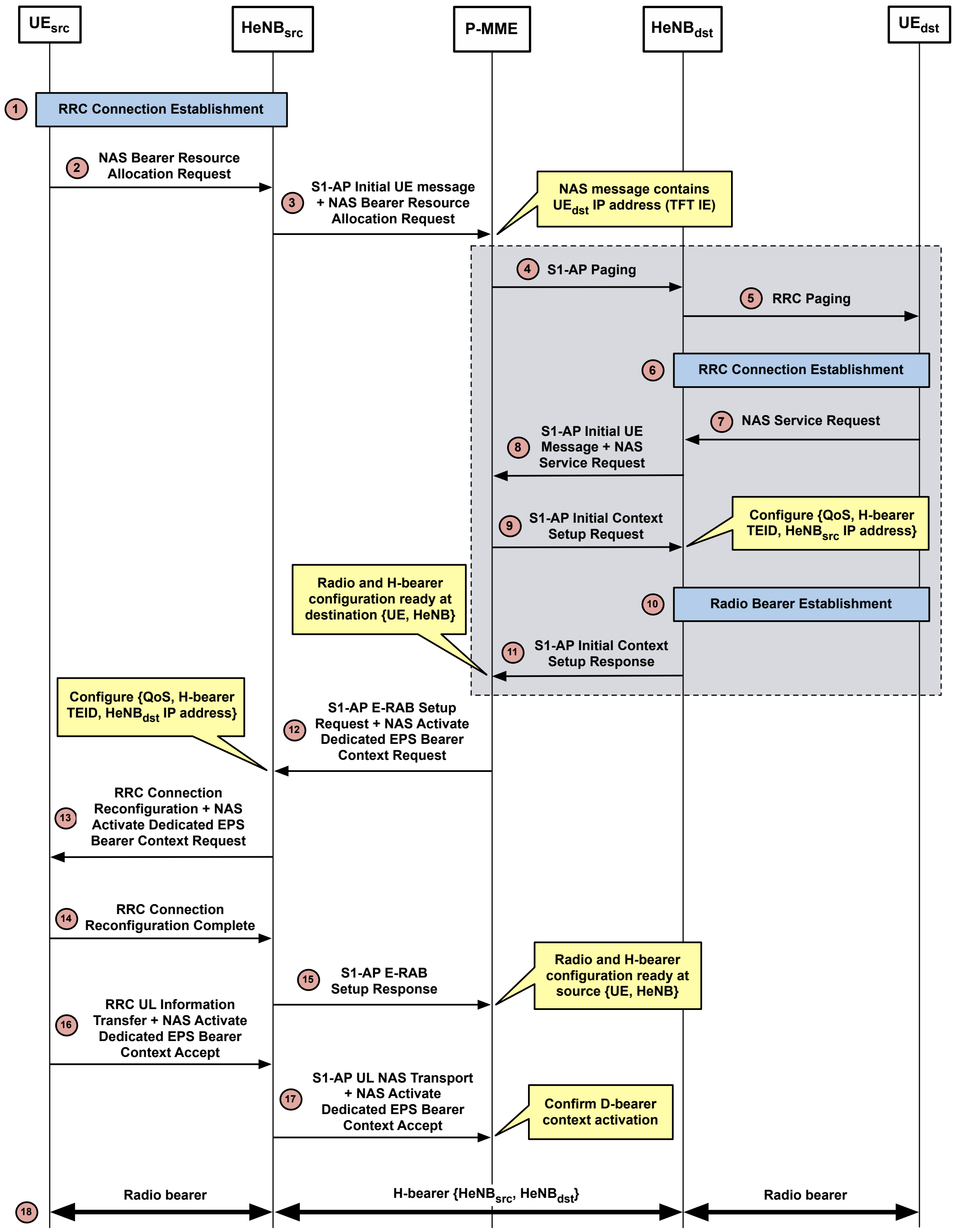


Figure 45.- The D-bearer establishment procedure

- **Step 7:** UE_{dst} sends a *NAS Service Request* message to $HeNB_{dst}$ in order to initiate the establishment of the corresponding radio and S1 bearers. Note that, instead of establishing an S1 bearer with the P-SGW, the P-MME will trigger the establishment of an H-bearer between the source and destination HeNBs.

- **Step 8:** $HeNB_{dst}$ forwards the *NAS Service Request* message to the P-MME encapsulated in an *S1-AP Initial UE* message. The purpose of this message is to instruct the P-MME to establish the corresponding radio and S1 bearers in response to the incoming paging procedure.

- **Step 9:** the P-MME detects UE_{dst}'s response to the incoming paging procedure and replies with an *S1-AP Initial Context Setup Request* message to $HeNB_{dst}$. Amongst other configuration parameters, this message contains the necessary QoS requirements, the H-bearer tunnel endpoint ID (TEID), and the $HeNB_{src}$'s IP address. Note that in a conventional EPS bearer establishment procedure, the (P)-MME would have replied with the necessary network configuration to establish an S1 bearer with the S-GW/P-SGW.

- **Step 10:** $HeNB_{dst}$ receives the configuration parameters from the P-MME and proceeds to establish a radio bearer with UE_{dst} and configure the H-bearer at the $HeNB_{dst}$ endpoint. Note that this operation follows the standard 3GPP procedure, except for the fact that the configuration settings provided by the P-MME correspond to the H-bearer instead of the S1 bearer.

- **Step 11:** $HeNB_{dst}$ sends an *S1-AP Initial Context Setup Response* message to notify the P-MME that the radio and H-bearer have been successfully configured at UE_{dst} and $HeNB_{dst}$ endpoints. Essentially, this marks the end of the network-initiated bearer establishment procedure initiated in step 4. Note that, at this point, a radio bearer has successfully been established between $HeNB_{dst}$ and UE_{dst}, and the H-bearer configuration has been applied at the $HeNB_{dst}$ endpoint.

- **Step 12:** after completing all the necessary steps at the destination end, the P-MME's attention shifts to UE_{src} and $HeNB_{src}$. This way, the P-MME sends a *NAS Activate Dedicated EPS Bearer Context Request* message encapsulated in an *S1-AP E-RAB Setup Request* message to $HeNB_{src}$ in order to (a) establish a dedicated user-plane radio bearer between UE_{src} and $HeNB_{src}$, (b) complete the H-bearer configuration at the $HeNB_{src}$ end, and (c) activate the D-bearer context between UE_{src} and UE_{dst}.

- **Step 13:** upon reception of the *S1-AP E-RAB Setup Request* message, $HeNB_{src}$ initiates a radio bearer establishment procedure with UE_{src} according to the requested QoS, H-bearer TEID, and $HeNB_{dst}$ IP address configuration parameters provided in step 12.

- **Steps 13 – 14:** $HeNB_{src}$ establishes a dedicated user-plane radio bearer with UE_{src}. Note that, at the $HeNB_{src}$ endpoint, this radio bearer is connected to the H-bearer terminated at $HeNB_{dst}$. UE_{src} indicates the successful establishment of the radio bearer by sending an *RRC Connection Reconfiguration Complete* message to $HeNB_{src}$.

- **Step 15:** upon reception of the *RRC Connection Reconfiguration Complete* message, $HeNB_{src}$ sends an *S1-AP E-RAB Setup Response* message to the P-MME. This message indicates that the radio and H-bearer configuration has successfully been applied at both the UE_{src} and $HeNB_{src}$ endpoints.

- **Steps 16 – 17:** finally, UE_{src} acknowledges the activation of the EPS bearer (in this case, the D-bearer) to the P-MME by sending a *NAS Activate Dedicated EPS Bearer Context Accept* message encapsulated in an *RRC UL Information Transfer* and *S1-AP UL NAS Transport* messages, respectively.

- **Step 18:** at the end of the D-bearer establishment procedure, the source and destination user-plane radio bearers between UE_{src} and $HeNB_{src}$, and UE_{dst} and $HeNB_{dst}$, respectively, have successfully been established. In addition, a H-bearer has also been established between $HeNB_{src}$ and $HeNB_{dst}$.

7.1.4 Handover Support for D-Bearers

Support of direct user-plane bearers between UEs during the execution of a handover procedure is a key traffic management feature in networks of small cells. In summary, this allows UEs engaged in direct communication to reallocate their already established D-bearer(s) endpoints during handover execution. This applies to both asymmetric and symmetric handover scenarios, i.e., where either one (or both) UEs engaged in a direct communication perform a handover. Figure 46 shows the message sequence chart of an asymmetric D-bearer handover where UE_{src} moves from $HeNB_1$ to $HeNB_2$. The exact procedure applies to analogous mobility scenarios (where UE_{dst} (or both UEs) perform a handover).

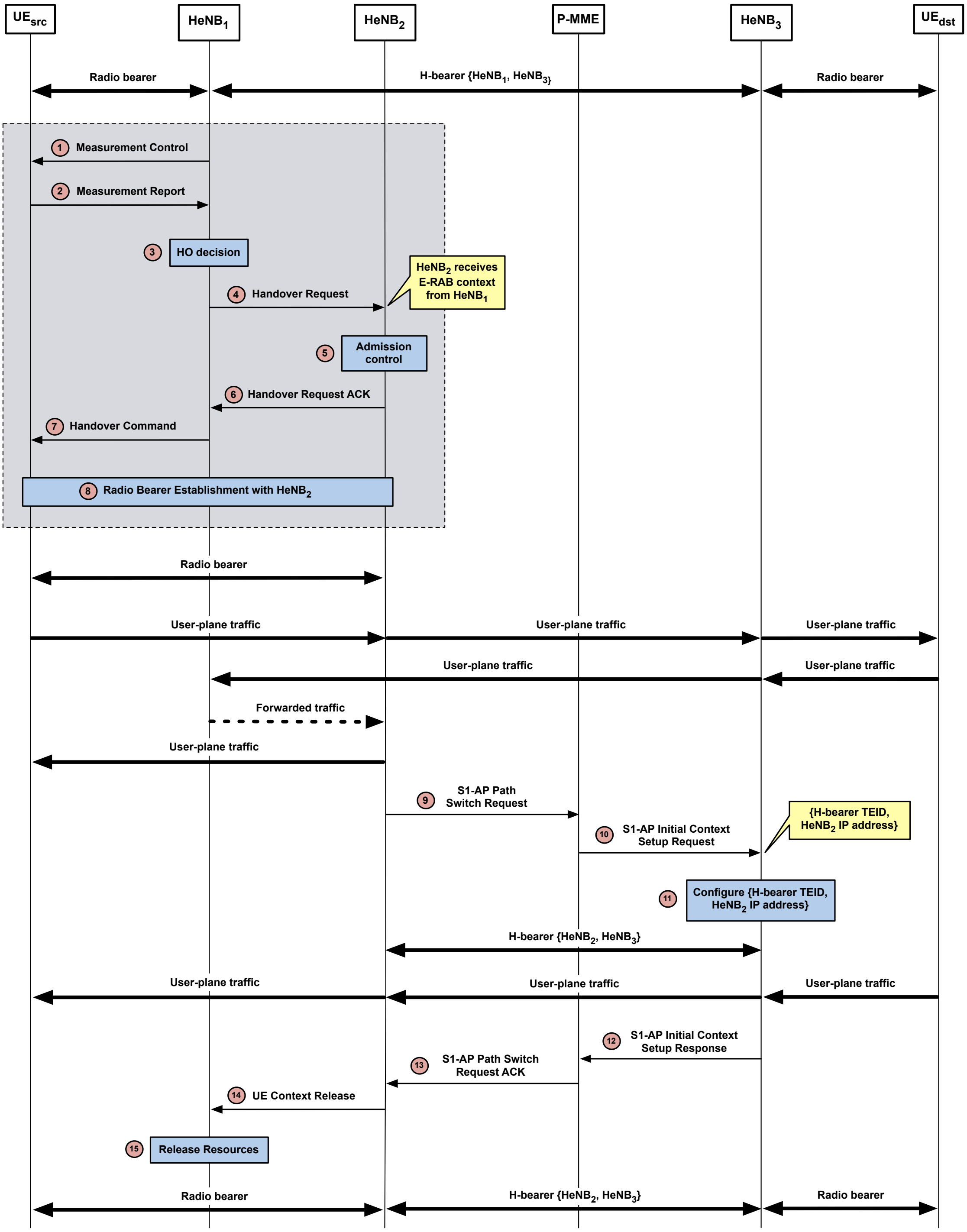


Figure 46.- Handover support for direct bearer communication

Prior to the initiation of the handover procedure, we assume that a D-bearer has already been established between UE_{src} (camped on $HeNB_1$) and UE_{dst} (camped on $HeNB_3$). Similarly to the direct bearer establishment procedure described in Section 7.1.3, the cornerstone of the D-bearer handover procedure is to use the *S1-AP Path Switch Request* message in a conventional handover (shown in the grey box in Figure

46) to trigger the reallocation of the H-bearer to the destination HeNB endpoint. A detailed description of the D-bearer handover procedure follows:

- **Steps 1 – 2:** $HeNB_1$ and UE_{src} exchange standard *RRC Measurement Control* and *RRC Measurement Report* messages to configure L1 measurements (and trigger the corresponding reports) at the UE_{src} end. As explained in Section 2.3.3 and [7], these measurements will be used by $HeNB_1$ to determine whether a handover procedure must be initiated (i.e., *handover decision*).

- **Step 3:** based on the measurement reports provided by UE_{src}, $HeNB_1$ initiates a handover procedure towards $HeNB_2$.

- **Step 4:** $HeNB_1$ sends an *X2-AP Handover Request* message to $HeNB_2$, along with the required configuration to prepare the handover at the target cell (i.e., the E-RAB context from $HeNB_1$). In sum, the aim of this message is to instruct $HeNB_2$ to perform admission control functions in order to determine if the necessary radio and network resources can be granted to the incoming UE.

- **Step 5:** $HeNB_2$ performs admission control based on its current available resources and the QoS requirements provided by $HeNB_1$. If these resources can be granted, $HeNB_2$ proceeds with the execution of the handover procedure. Otherwise, the handover is cancelled.

- **Step 6:** $HeNB_2$ configures the corresponding L1/L2 resources in its protocol stack and sends the *X2-AP Handover Request Acknowledge* message to $HeNB_1$. In addition to the necessary configuration parameters, this message also encapsulates a transparent container that will later be sent as an RRC message to UE_{src} to perform the handover (step 7).

- **Step 7:** $HeNB_1$ decapsulates the transparent container received from $HeNB_2$ and forwards the handover command message (*RRC Connection Reconfiguration*) to UE_{src}. Upon reception of this message, UE_{src} detaches from $HeNB_1$ and attempts to synchronise with $HeNB_2$. As shown in Figure 46, UE_{src} establishes a dedicated user-plane radio bearer with $HeNB_2$ in order to gain access to the E-UTRAN via the new HeNB. Note that, at this point, the *forward-directed* traffic (i.e., source to destination UE) is sent from UE_{src} to $HeNB_2$ over the newly allocated radio bearer, then from $HeNB_2$ to $HeNB_3$ (via the original H-bearer), and finally from $HeNB_3$ to UE_{dst} (via a previously allocated radio bearer). Instead, the *backward-directed* traffic (destination to source UE) is sent from UE_{dst} to $HeNB_3$ (via a previously allocated radio bearer), from $HeNB_3$ to $HeNB_1$ (via the original H-bearer), from $HeNB_1$ to $HeNB_2$ (via the X2 interface), and, finally, from $HeNB_2$ to UE_{src} over the newly allocated radio bearer. The remaining steps in the D-bearer handover procedure focus on reallocating H-bearer's source endpoint from $HeNB_1$ to $HeNB_2$.

- **Step 9:** $HeNB_2$ sends an *S1-AP Path Switch Request* message to the P-MME. In a conventional handover scenario, this message would trigger the necessary EPC procedures to enable user-plane traffic redirection from the S-GW to the new cell, and vice-versa. However, in a D-bearer handover, the P-MME uses this message to instruct $HeNB_3$ that the H-bearer's source endpoint has been reallocated from $HeNB_1$ to $HeNB_2$.

- **Step 10:** the P-MME sends an *S1-AP Initial Context Setup Request* message to apply the new H-bearer configuration in $HeNB_3$. Amongst other parameters, this message contains the new H-bearer tunnel endpoint ID, as well as $HeNB_2$'s IP address.

- **Step 11:** $HeNB_3$ applies the new H-bearer configuration locally, thus completing the H-bearer establishment between $HeNB_2$ and $HeNB_3$. From this point onwards, backward-directed traffic (destination to source UE) will be sent from UE_{dst} to $HeNB_3$ (over the previously allocated radio bearer), from $HeNB_3$ to $HeNB_2$ (over the newly allocated H-bearer), and from $HeNB_2$ to UE_{src} (over the newly allocated radio bearer).

- **Step 12:** $HeNB_3$ responds to the *S1-AP Initial Context Setup Request* message sent in step 10 with an *S1-AP Initial Context Setup Response* message.

- **Step 13:** the P-MME responds to the *S1-AP Path Switch Request* message in step 9 with an *S1-AP Path Switch Request Acknowledge* message. The purpose of this message is twofold: on the one hand, it confirms the successful reallocation of the H-bearer at the destination endpoint of the D-bearer. On the other hand, it instructs $HeNB_2$ to inform $HeNB_1$ that the old radio bearer between UE_{src} and $HeNB_1$ (along with its associated network resources) can now safely be released.

- **Steps 14 – 15:** $HeNB_2$ instructs $HeNB_1$ to release the old radio bearer between UE_{src} and $HeNB_1$, along with any associated network resources. This operation is performed in step 15. Note that, once the D-bearer handover procedure has been completed, the resulting direct bearer comprises a dedicated user-plane radio bearer between UE_{src} and $HeNB_2$, a newly allocated H-bearer between $HeNB_2$ and $HeNB_3$, and a dedicated user-plane radio bearer between $HeNB_3$ and UE_{dst}.

7.2 Multicast Traffic Management

The use of direct bearers between UEs, along with the deployment of distributed routing protocols in the transport network layer, addresses the issue of unicast traffic management in large-scale, all-wireless networks of small cells. However, support for efficient multicast/broadcast services in these scenarios still remains an open issue.

As explained in Section 2.4.2, the Multicast/Broadcast Multimedia Service (MBMS) is the standard 3GPP mechanism to support multicast and broadcast traffic distribution in the Evolved Packet System. MBMS for UMTS was first introduced in Release 6 of 3GPP Technical Specifications. Instead, the Evolved Packet System version of MBMS was initially published in Release 9 of 3GPP Technical Specifications. Some popular MBMS use cases comprise live video and audio streaming, push media, e-publications, OS updates, newsreel information, weather forecasts, traffic updates, etc.

In terms of network performance, MBMS suffers from the same limitations of conventional multicast/broadcast wireless systems, i.e., the shortage of radio spectrum (only up to 60% of LTE subframes can be allocated to MBMS traffic [103]) and low multicast/broadcast transmission bit rate to ensure proper reception (and, thus, QoE) for cell-edge users. Large-scale deployments of small cells specifically aimed at supporting multicast and broadcast services in localised areas can help to mitigate these issues. However, these deployments do not address the intrinsic fixed bit rate limitations of MBMS caused by the lack of feedback mechanisms between UEs and

(H)eNBs. In this chapter we propose a RAT-agnostic Adaptive Multicast Service (AMuSe) aimed at overcoming the traditional limitations of wireless multicast distribution systems; in particular, those related with the lack of feedback mechanisms capable of adapting the Modulation and Coding Scheme (MCS) and, thus, the transmission bit rate, to the available network and radio resources. Our proposal supports scalable and efficient delivery of wireless multicast/broadcast traffic to a large number of end users with low communication overhead. In particular, AMuSe enables the transmitting nodes to send multicast traffic at higher bit rates whilst ensuring high Packet Delivery Ratio (PDR) for the vast majority of the receiver nodes[7] (over 95%).

The AMuSe system has been specifically designed to be RAT-agnostic, i.e., it can be deployed indistinctively in licensed (e.g., UMTS, LTE, LTE-A) and unlicensed (e.g., Wi-Fi) wireless systems as a standalone application-layer function in the control-plane protocol stack. The reasons leading to this design strategy are twofold: on the one hand, encapsulating all AMuSe functionalities in a standalone application-layer entity in the control-plane protocol stack facilitates its implementation in commercial network equipment, namely (H)eNBs and UEs (in the case of 3GPP systems) and access points (APs) and 802.11 clients (for Wi-Fi systems). This way, AMuSe's system functions can be deployed in commercial devices as vendor-specific features with small or no impact on the standard LTE/802.11 protocol stack architecture and network procedures. The second reason stems from the practical limitations of carrying out an extensive experimental evaluation of adaptive multicast/broadcast traffic distribution services in a large-scale LTE testbed. In order to control the MCS settings and the transmission bit rate of a (H)eNB, upper-layer entities in the control-plane protocol stack need to interact with the RLC and MAC layers in the LTE-Uu interface. However, commercial (H)eNBs are purpose-specific cellular network devices where lower-layer service primitives providing control over radio resource management functions are usually not exposed to upper layers in the protocol stack. Although some emerging LTE testbed platforms have recently gained attention in the academic community [104], [105], [106] at the time of writing none of

[7] We assume that the multicast/broadcast flow contains some redundancy (e.g., FEC) across packets in order to deal with a limited amount of losses.

them supported an open-access, large-scale implementation of multicast/broadcast traffic distribution systems. In consequence, we decided to decouple AMuSe's core feedback selection and rate adaptation functions from the underlying radio access technology (e.g., LTE, 802.11abgn, etc.) in order to facilitate its functional evaluation in a general-purpose wireless experimental testbed. In particular, we selected the ORBIT testbed [107] located in the WINLAB research facilities at Rutgers University, New Jersey (United States), due to its large-scale dimension (400 wireless nodes) and widespread acceptance in the academic community. In addition, all ORBIT nodes deploy Application Programming Interfaces (APIs) that provide full access to the underlying MAC and PHY configuration in the protocol stack. For further information on the ORBIT testbed the reader is referred to [107] and [108].

The following sections describe the research work carried out during the design, development, and experimental evaluation of the AMuSe system in the ORBIT testbed. Section 7.2.1 provides a detailed description of AMuSe's feedback selection mechanism, as well as a proposal for the control-plane protocol between the sender and receiver nodes. Section 7.2.2 describes the implementation and experimental evaluation of the AMuSe system in the ORBIT testbed. Finally, Section 7.2.3 discusses the implications of the AMuSe system in the MBMS network and protocol stack architectures.

7.2.1 The AMuSe System

The AMuSe system is based on the observation that a cluster of adjacent wireless nodes experience similar channel quality and interference patterns [109]. AMuSe dynamically divides the entire receiver population into groups of nodes or *clusters* based on distance between nodes. Then, a single node in each cluster is selected to report QoS measurements to the AP (e.g., channel quality, PDR, lost packets, etc.). This node is referred to as the *feedback node* (FB node). The AP, in response, may decide to report back to the multicast source node, adjust the FEC mechanism, adapt the modulation and coding scheme (and, thus, the transmission bit rate), retransmit lost packets, etc.

AMuSe supports a simple and efficient FB node selection process that requires a small number of nodes to send feedback messages to the AP. This lightweight mechanism results in low communication overhead over the wireless channel. The process strikes a balance between the number of FB nodes, the accuracy of the feedback reports, and the system convergence time by controlling the maximal radius of the clusters, denoted by D. The feedback selection mechanism ensures that every node is, at most, within a D distance from another FB node that is experiencing similar or weaker radio channel conditions. Thus, in order to ensure a sparse density of feedback nodes, any pair of FB nodes must be, at least, D distance units apart form each other. This is a variant of the well-known minimal independent dominating set problem [110]. Although this problem is known to be NP-hard, our distributed algorithm can find a solution with small constant approximation ratio when paired with a widely used signal propagation model. To this extent, AMuSe offers the following salient features:

- **RAT-agnostic:** the proposed scheme can be deployed in any wireless device (e.g., 802.11, LTE, etc.) as an application-layer entity in the control-plane protocol stack. This applies to both network equipment (e.g., Wi-Fi access points, (H)eNBs) and mobile devices (802.11 clients, UEs).
- **Efficient wireless resource utilisation:** AMuSe selects a small number of FB nodes efficiently to produce a limited amount of communication overhead whilst accurately reflecting the PDRs experienced by wireless nodes.
- **Scalable solution:** cluster partition depends only on the cluster radius (D) instead of node density. Thus, the AMuSe system scales as a function of the geographical area.

As previously explained, we have used the large-scale, general-purpose ORBIT testbed to carry out the experimental evaluation of the AMuSe system. During the course of our experiments we noticed that some nodes (referred to as *abnormal nodes*) suffered from low PDR even when the AP was using low transmission bit rates (i.e., robust MCS) and even in the absence of external interference. Furthermore, this set of abnormal nodes varied across different experiment repetitions. Abnormal nodes pose a considerable challenge in feedback selection and rate adaptation mechanisms, as they cause low network utilisation even when most of the non-abnormal nodes could benefit from much higher transmission bit rates. In order to address this issue, we designed the AMuSe system in a way such that abnormal nodes could be detected and, eventually, isolated from normal FB nodes. We showed that, in combination with a simple MCS selection/rate adaptation mechanism running in the AP, the AMuSe system was able to improve system throughput while ensuring a PDR of 90% for more than 95% of the nodes. This is described in detail in Section 7.2.2.

7.2.1.1 Network Scenario

Our network scenario considers an IEEE 802.11 wireless local area network (WLAN), with multiple access points that serve a very dense population of mobile users. Due to the shortage of orthogonal wireless channels, adjacent APs may use the same channels. In consequence, a transmission in one cell may interfere with transmissions in its neighbouring cells[8]. In general, we consider transmissions from both mobile devices and APs outside of each cell as potential sources of interference. In particular, we focus on a single cell with a given AP. In this scenario, the AP sends a multicast traffic flow to a large group of mobile nodes in its transmission range.

As described in the previous section, we follow a model where a node may send reports to an AP (or, alternatively, a multicast server) about its experienced quality of service (e.g., link quality, PDR, lost packets, etc.). In turn, the AP/multicast server may decide to adjust the FEC mechanism, adapt the MCS/transmission bit rate, retransmit lost packets, etc. In practice, the AP and the multicast server are two separate functional entities. These functions may reside in the AP or the multicast server and may also be distributed across multiple layers in the protocol stack. However, functions in the AP are the only ones responsible for adjusting transmission parameters. Thus, in the following sections we will use the term AP in a broad sense to denote a combination of the AP and multicast server functions.

At any given time each wireless node is associated with a single AP. In addition, nodes are assumed to have a quasi-static mobility pattern, i.e., they are free to move

---

[8] Note that this network scenario, whilst being characteristic from Wi-Fi systems, is also applicable to large-scale deployments of networks of small cells.

from place to place, although they tend to stay in the same physical location for several minutes. This is a reasonable assumption for crowded scenarios, such as sports venues and transportation hubs. We assume that mobile devices can estimate their approximate locations by using one of the methods described in [111] or any other client/network-based indoor or outdoor positioning system.

7.2.1.2 Design Objectives

The AMuSe system aims at supporting the efficient transmission of wireless multicast/broadcast user-plane traffic at high bit rates with minimal overhead, whilst ensuring a high packet delivery ratio for the majority of nodes. In particular, AMuSe must ensure that at least X% of the nodes (e.g., X = 95) experience PDR values above a given threshold denoted by H (e.g., H = 90%).

In order to realise a system design that fulfils the previous design objectives, we rely on the observation reported in related studies [109], i.e., *a cluster of adjacent wireless nodes experiences similar channel conditions and suffers from similar interferences. Hence, a node v with worse channel conditions than its adjacent neighbours can effectively represent the QoS observed by the rest of the nodes in the cluster. Furthermore, if a packet is correctly received by v then, with high probability, the rest of the nodes in the cluster will also receive it.*

Based on this observation, we divide all nodes in the wireless network into clusters of adjacent nodes and then select a single feedback node (FB node) to represent the remaining nodes in the cluster (i.e., to report QoS metrics on multicast traffic reception to the corresponding AP). In turn, the AP will use these reports to improve the quality of the multicast service, e.g., by adding FEC, adjusting the MCS, adapting the transmission bit rate, retransmitting lost packets, etc.

Rather than focusing on the necessary actions for improving the quality of the multicast transmission on the AP side, we primarily focus on designing an efficient feedback selection mechanism to determine the set of FB nodes amongst all wireless receivers. Thus, the proposed feedback mechanism aims at finding a small set of well-distributed FB nodes that can provide QoS reports according to a given level of accuracy.

7.2.1.3 Feedback Selection Mechanism

Two wireless nodes are defined to be *D-adjacent* if the physical distance between them is, at most, D. In order to find a small set of FB nodes to provide accurate QoS reports, all wireless nodes in the system must satisfy the following requirements:

(a) Each node must be D-adjacent to a FB node.

(b) A FB node must have similar or lower QoS than any of its D-adjacent nodes.

(c) Two FB-nodes cannot be D-adjacent to each other.

With regards to the feedback selection mechanism, we propose a semi-distributed scheme for selecting FB nodes where the AP selects the best candidates from a set of wireless nodes that volunteer to serve as FB nodes (*volunteer nodes*). If the approximate node location (as well as their observed link quality) is known, the AP can easily select the ideal set of FB nodes. However, this might not be feasible for large groups of wireless receivers. Thus, we seek to minimise the number of nodes that send periodical QoS reports to the AP as part of the FB node selection process whilst ensuring that a small set of FB nodes meet the above requirements.

The AP periodically sends a feedback list message (FBN-LIST) to all wireless nodes in the network. This message contains the list of all FB nodes in the multicast session. In the experimental evaluation of the AMuSe system, the FBN-LIST message is sent every τ_{AP} = 500 ms. As shown in Figure 47, each entry in the FBN-LIST contains a temporary node ID, its reported location (physical/virtual (X,Y) coordinates), its reported link quality, and a packet delivery ratio estimate.

Node ID	(X,Y)	Channel Quality	PDR

Figure 47.- Structure of a feedback list message entry

In the AMuSe system, a wireless node is always in one of the following states:

- **FB-NODE:** a node in FB-NODE state has been selected to become a FB node in its own cluster.

- **VOLUNTEER:** a node in VOLUNTEER state is not aware of any D-adjacent FB nodes that are experiencing lower or similar QoS conditions. Thus, a node in VOLUNTEER state is eligible to serve as a FB node in its own cluster.
- **NON-FB-NODE:** a node in NON-FB-NODE state is aware of a D-adjacent FB node that experiences a similar or lower QoS conditions.

Figure 48 depicts the state transition diagram of AMuSe's feedback selection mechanism. As shown in the figure, when a node v joins the system it is automatically put into VOLUNTEER state. Then, it waits for the arrival of an FBN-LIST message and checks if there are any D-adjacent FB nodes in the feedback list with similar or worse QoS conditions. If there are any such nodes, v moves into NON-FB-NODE state and records the list of D-adjacent FB nodes in the FBN-LIST message with similar or worse QoS conditions. If there are no such nodes, v starts a random back-off timer for a period in the interval $[0, T]$ (in our experiments we set the maximum receiver back-off timer, T, to 5 seconds). During this countdown, if v learns of a D-adjacent FB node by inspecting the periodic FBN-LIST messages it stops its countdown and moves into NON-FB-NODE state. Otherwise, upon expiration of the timer, it sends an FBN-JOIN message to the AP and waits to see if its node ID appears on the next FBN-LIST (as described in Figure 47, the FBN-JOIN message contains the node ID, the node location, observed link quality, and the estimated PDR). If v appears on the FBN-LIST it moves into FB-NODE state. Otherwise, it repeats the back-off process until it leaves the VOLUNTEER state. At any time, upon reception of an FBN-LIST message, if a FB node v does not find its node ID on the feedback list it leaves the FB-NODE state. If a D-adjacent FB node appears on the list, it moves into NON-FB-NODE state; otherwise, it switches to VOLUNTEER state.

An important property of AMuSe's volunteering process is that FB node selection is done in a semi-distributed manner, i.e., a wireless node volunteers to serve as a FB node if (and only if) if there are no other FB nodes in its vicinity that experience worse QoS conditions. Thus, it is the responsibility of the AP to resolve potential conflicts when several D-adjacent nodes volunteer simultaneously to become FB nodes, as well as to remove unnecessary FB nodes from the feedback list.

Consequently, before sending the FBN-LIST message to all nodes, the AP needs to execute a node-pruning algorithm to determine which nodes will effectively serve as FB nodes and which will be discarded. This algorithm is described in Section 7.2.1.4.

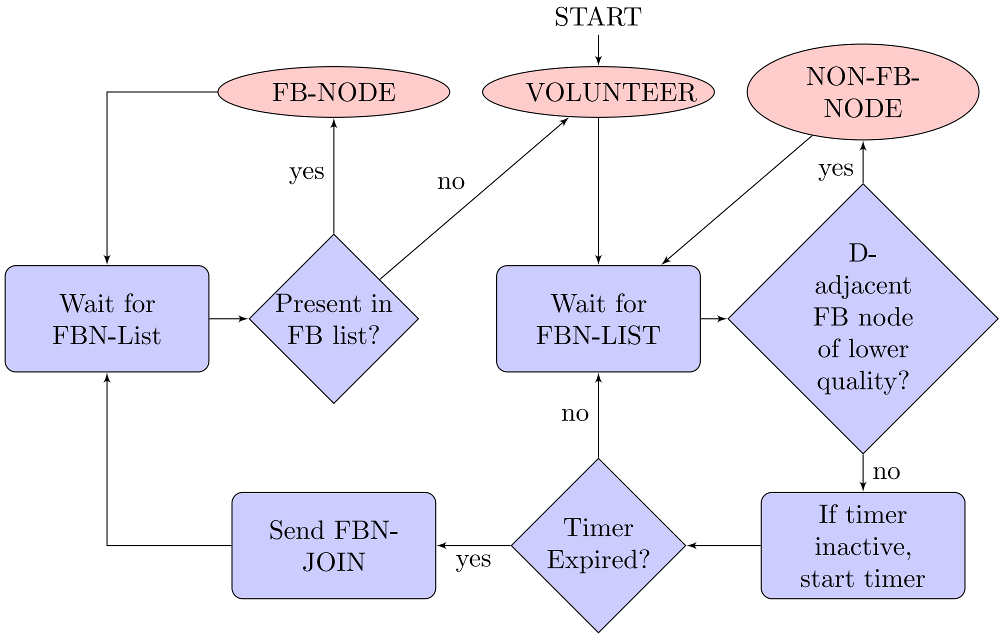


Figure 48.- State transition diagram of AMuSe's feedback selection mechanism

Each FB node periodically sends REPORT messages to inform the AP about the current channel and multicast service quality that the FB node (and, thus, the cluster it represents) is experiencing. In the experimental evaluation of the AMuSe system, we set this value to $\tau_{FB} = 500$ ms. If the AP does not receive any REPORT messages from one of the FB nodes in the feedback list for a certain period (e.g., $3\tau_{FB}$ in our experiments), the AP removes the FB node from the feedback list.

Some aspects of AMuSe's feedback selection mechanism are worth pointing out here. First, we do not require the nodes to listen to all the traffic in the network. Instead, they just need to listen to AP transmissions on the multicast group address, thus saving energy at the receivers. Second, we do not require wireless nodes to submit very precise location information. As previously explained, coarse granularity is acceptable as long as the accuracy is in the order of a few meters. Third, we provide variable levels of reliability by fine-tuning the combination of AP reporting frequency (τ_{AP}), receiver reporting frequency (τ_{FB}), maximum receiver back-off timer (T), and node adjacency distance (D), with more reliable and frequent reports leading to more signalling overhead in the wireless channel. This way, instead of providing a setup based on a single configuration parameter, we support multiple control knobs, thus allowing a greater flexibility to the mobile network operator depending on the nature of the multicast streams.

In order to illustrate the operation of AMuSe's feedback selection mechanism we consider the network topology depicted in Figure 49(a). As shown in the figure, this topology comprises a single AP (located in the bottom right corner of the figure) and four wireless receivers, each one located at a different distance from the AP. We assume that the label in each node denotes the node ID, as well as the temporal order in which it joins the multicast service at this AP. We consider four different link quality levels based on the distance between the AP and the nodes, namely *very good*, *good*, *fair*, and *poor* (these levels correspond to the wireless channel quality experienced by nodes 1, 2, 3, and 4, respectively). Figure 49(b) shows a circle of radius D centred at each node. According to the definition of *D-adjacency*, each node u contained in the circle centred at node v is D-adjacent to v. Thus, if v and u are D-adjacent, they are considered *neighbours* to each other.

In the following example we show the importance of the three requirements of AMuSe's feedback selection mechanism in the quality and density of the FB node set. First, let us assume that the FB nodes are only required to meet conditions (a) and (b), but not (c). Under this assumption, each time a node joins the multicast group it has a weaker channel quality than any of its neighbours and, therefore, it is selected as a FB node. Once the last node (ID = 4) has joined the multicast group the system contains exactly four FB nodes. It is straightforward to realise that this approach will not scale for large groups.

Now let us assume that requirement (c) is enforced. Right after a node joins the multicast group, the set of FB nodes is optimised. When node 1 joins, it becomes the first FB node. After node 2 joins, node 2 becomes the FB node, whilst node 1 becomes a non-FB node due to the enforcement of condition (c). After node 3 joins, it becomes a FB node whilst both node 1 and 2 become non-FB nodes because all three nodes are D-adjacent to each other. When node 4 joins it becomes a FB node, whilst node 3 becomes a non-FB node. In addition, node 2 becomes a FB node and node 1 remains as a non-FB node. Notice that, in this example, node 2 changes its state twice; first when node 3 joins the multicast group and, second, after node 4 joins the system. Notice also how the set of FB nodes is optimised after each node joins the multicast group.

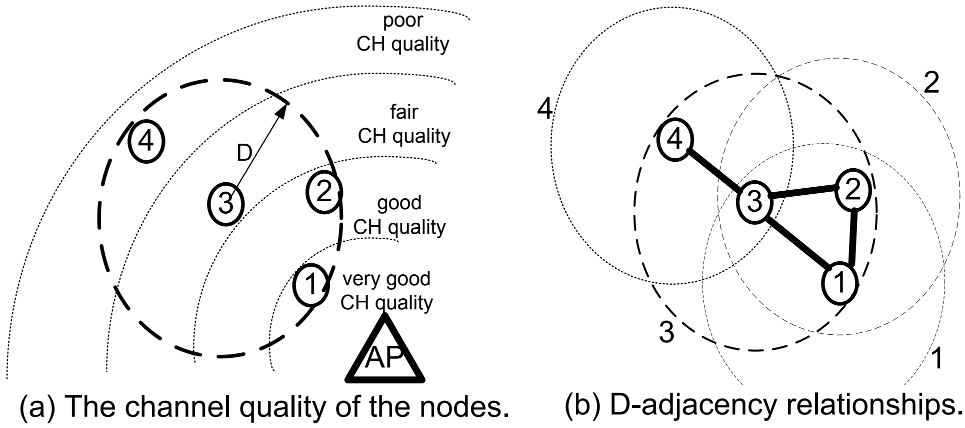


Figure 49.- Feedback selection example with 4 wireless receivers

In summary, this example shows that whilst a feedback selection algorithm that satisfies all three requirements may cause churn as nodes join and leave the FB node set, the selected set of FB nodes is near optimal when the set of nodes in the multicast group does not change.

7.2.1.4 The Node-Pruning Algorithm

As previously explained, AMuSe's feedback selection mechanism ensures that every node is D-adjacent to a candidate node with similar or worse QoS conditions, where the term *candidate node* refers to current FB nodes as well as those in VOLUNTEER state. Thus, it is the responsibility of the AP to discard unnecessary candidates in order to define a small set of FB nodes such that any pair of feedback nodes are not D-adjacent to each other. This is known as the *node-pruning algorithm*.

The problem of finding the minimal set of FB nodes that meets the three requirements stated in Section 7.2.1.3 is a variant of the minimal dominating-set problem, which is a well-known NP-complete problem also in the case of unit disk graph [110]. Below we present a heuristic algorithm for selecting a near-optimal set of candidate nodes that meets the three requirements previously stated:

1. The AP creates a list L of the candidate nodes, i.e., containing all *feedback* and *volunteer* nodes.
2. The AP sorts the candidate list in increasing order according to each node's QoS condition (channel quality, PDR, LQ, etc.).
3. The AP iteratively selects the first candidate v in L as a FB node and removes v (along with all its D-adjacent nodes) from L.
4. The AP returns to step 3 until L is empty.

A naive implementation of the heuristic algorithm may cause FB node churn, which, in turn, could have an impact on system stability. Since node pruning is performed at the access point level, the algorithm can be easily modified to prevent node churn, e.g., by giving higher priorities to already selected FB nodes or, instead, by relaxing the distance constraint between FB nodes. The in-depth discussion about striking a proper balance between system stability and output optimality in system design lies beyond the scope of this Ph.D. Thesis.

7.2.2 Experimental Evaluation

In this section we provide a detailed description of the experimental evaluation of the AMuSe system in the framework of the ORBIT testbed. First, we describe the software architecture to support the implementation of the distributed feedback selection and rate adaptation mechanisms in the ORBIT nodes. Secondly, we present the *offline* experiments carried out to evaluate the performance of AMuSe's feedback selection mechanism. In particular, this comprises the experimental collection of large amounts of radio channel and QoS measurements, as well as the offline post-processing of the obtained results with MATLAB. Finally, we discuss the *online* experimental evaluation of the AMuSe system, i.e., the real-time deployment of AMuSe's feedback selection and rate adaptation mechanisms in the ORBIT testbed in order to illustrate how the AMuSe system can effectively achieve higher system throughput whilst ensuring high PDR values for the majority of the wireless nodes in the network.

7.2.2.1 *Software Architecture*

Figure 50 shows the software modules, protocol messages, network ports, and socket interfaces used to implement AMuSe's distributed feedback selection and rate adaptation mechanisms in the ORBIT testbed.

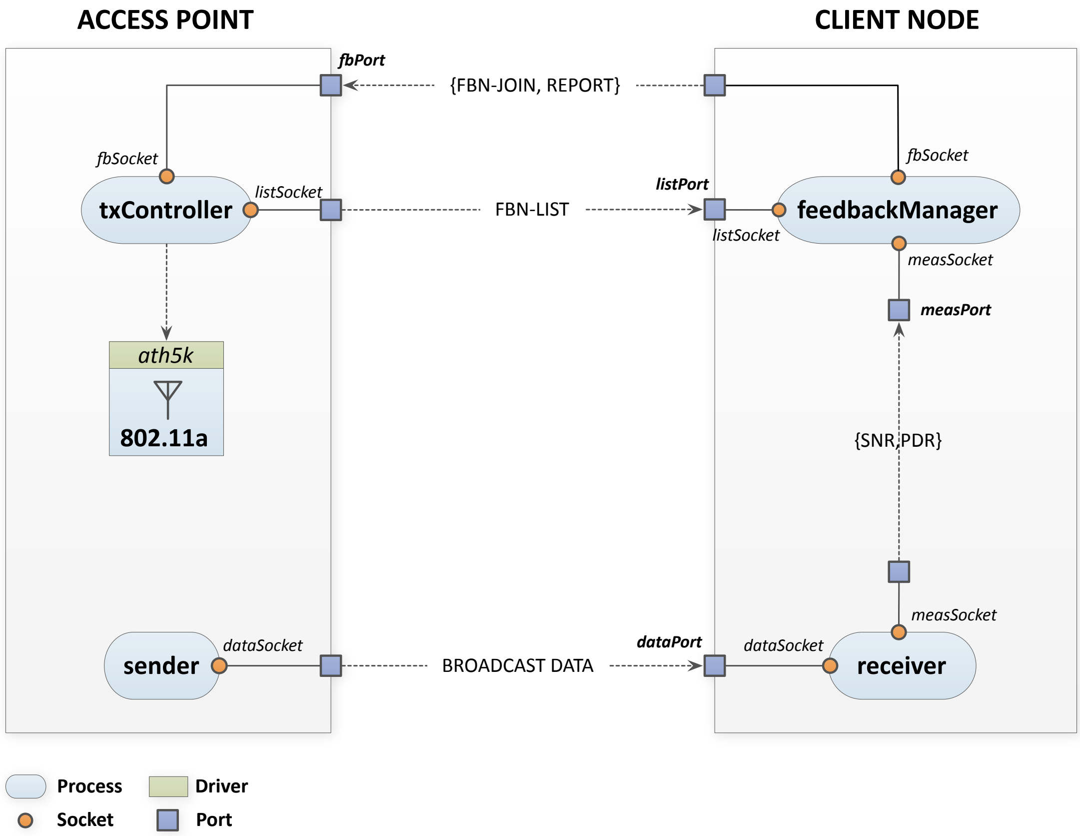


Figure 50.- Software architecture of the AMuSe system

As shown in the figure, the software architecture is split into two generic nodes, namely the *access point* (left) and the *client* (right). On the one hand, the access point hosts two concurrent software modules: the `sender` and the `txController`. On the other hand, each wireless client in the ORBIT testbed hosts concurrent instances of the `receiver` and `feedbackManager` modules. Each software module implements a different set of AMuSe's system functions, as described below:

- **`sender`:** the `sender` module in the AP is responsible for sending broadcast traffic to all wireless clients in the network. In order to emulate high-traffic load conditions, this module generates a constant flow of 1400-byte long application messages encapsulated into UDP datagrams. In turn, the AP encapsulates these messages into 802.11a frames that are sent to all client nodes in a broadcast fashion. By setting the length of the application message to 1400 bytes we achieve high network load conditions without triggering fragmentation mechanisms in the AP MAC layer.

- **`receiver`**: the `receiver` module is the natural counterpart to the `sender` module in the AP. Instances of the `receiver` module run in each client in the ORBIT testbed acting as a traffic sink for the broadcast 802.11a frames sent by the `sender`. The `receiver` module consists of a UDP server listening on the `dataPort` port to UDP datagrams sent by the `sender` module, as well as of a UDP client that periodically sends QoS measurements (SNR and PDR) to its corresponding local `feedbackManager` module. The `receiver` module obtains SNR measurements by periodically polling the local Atheros 5212/5213 wireless cards via standard ath5k wireless driver commands. Analogously, the `receiver` module computes PDR estimates by using a custom sequence number embedded in the application-layer payload sent by the `sender` module.

- **`feedbackManager`**: similarly to the `receiver` module, the `feedbackManager` module implements the client side of the feedback selection and rate adaptation algorithm intelligence in all client nodes. In practice, it consists of a UDP server receiving periodic QoS measurements from the local `receiver` module on the `measPort` port, as well as of a UDP server listening to periodic FBN-LIST messages on the `listPort` port sent by the `txController` module in the AP. Furthermore, it also implements a UDP client that sends FBN-JOIN and REPORT messages to the `txController` module in the access point. In sum, the `feedbackManager` module analyses QoS measurements sent by the local `receiver` module (in conjunction with the information contained in the FBN-LIST messages sent by the access point) and determines whether a client node must volunteer to become a feedback node (FBN-JOIN), report QoS measurements to the access point (REPORT), or remain silent.

- **`txController`**: the `txController` module implements the AP side of the feedback selection and rate adaptation algorithm intelligence. In particular, the `txController` module consists of a UDP server listening to REPORT and

FBN-JOIN messages from FB nodes on the `fbPort` port in order to update the contents of the feedback list. In addition, it periodically sends the FBN-LIST message to all client nodes via a local UDP client. The `txController` is also in charge of adapting the transmission bit rate (and, hence, the modulation and coding scheme) in the AP PHY layer by interacting with the local 802.11a wireless card through the ath5k wireless driver.

With regards to system implementation, all software modules have been developed in C and run as concurrent user-space processes in the ORBIT nodes. These processes communicate with the ath5k wireless drivers in each node via standard Linux shell commands, such as `iwconfig`. All wireless nodes in the online (and offline) experiments are equipped with Atheros 5212/5213 series wireless cards that expose their Application Programming Interface (API) to a Linux-based operating system (Ubuntu 12.04) via the ath5k driver. In order to enable variable transmission bit rate selection for broadcast user-plane frames in conjunction with fixed transmission bit rate (6 Mbps) for 802.11a beacon frames, we developed a custom version of the ath5k driver for the AP node.

7.2.2.2 Offline Experiments

In this section we describe the offline experimental campaign carried out in order to evaluate the performance of AMuSe's feedback selection mechanism. A full-fledged evaluation of the AMuSe system (i.e., feedback selection and rate adaptation) is presented in Section 7.2.2.3. During the offline campaign we ran multiple experiment repetitions in order to collect large amounts of link quality and QoS data at each wireless node in the ORBIT testbed. Later on, we used this data to feed software implementations of various feedback selection algorithms in MATLAB in order to compare the performance of AMuSe's feedback mechanism against other schemes. In terms of QoS metric, we used the Link Quality (LQ) parameter[9] reported by the wireless cards in each client node as an estimate of its observed received signal strength (RSS). Prior to describing the obtained results, we first consider a set of

---

[9] Although LQ is not a standard radio channel quality metric, we observed that the LQ values reported by the Atheros 5212/5213 series chipsets indicated the RSS (in dB) normalised to a reference value of -110dBm (the constant thermal noise floor).

hypotheses used to validate the observation in Section 7.2.1.2 and thereby, the proposed system design.

- **Hypothesis 1 (H1):** there is some correlation between the Packet Delivery Ratio (PDR) and Link Quality (LQ) values observed by a wireless node.
- **Hypothesis 2 (H2):** clustered nodes experience similar LQ and PDR values.
- **Hypothesis 3 (H3):** clustered nodes suffer from similar interference.
- **Hypothesis 4 (H4):** a node with lower LQ than its neighbours may serve as a good FB node to represent the PDR observed by its neighbours.

As previously explained, the ORBIT testbed features a general-purpose, fully configurable grid of 20×20 (400) nodes equipped with 802.11abgn wireless cards. The physical separation between adjacent nodes is 1 meter. Furthermore, the testbed provides a noise generator with 4 noise antennas at the corners of the grid whose attenuation can be controlled independently. This allows researchers to emulate more complex network scenarios (e.g., external noise, shadowing, interference, etc.). In order to avoid performance issues stemming from a mismatch of 802.11 hardware and software, we selected a subset of ORBIT nodes equipped with Atheros 5212/5213 wireless cards in conjunction with ath5k wireless drivers[10]. In addition, we removed unresponsive nodes from the ORBIT grid (e.g., due to OS problems) before running each experiment, which resulted in approximately 250-300 active nodes per experiment repetition.

With regards to the experimental setup, the top-left node in the ORBIT grid (i.e., the node with virtual coordinates (1,1)) acted as a multicast AP (thus configured in 802.11 *master* mode) and used 802.11a channel 40[11] to send a multicast UDP flow with a transmission power of 1 mW = 0 dBm. The remaining wireless nodes in the ORBIT testbed acted as multicast receivers (thus configured in 802.11 *managed*

---

[10] Alternative available options comprised other 802.11 chipsets (e.g., Intel, Broadcom, NetGear, etc.) and 802.11 wireless drivers (ath9k, madwifi, etc.).

[11] We observed that channel 40 in the 5 GHz band suffered from less external interferences in the ORBIT grid than other channels in the 2.4GHz band.

mode). Every node kept a record of the parameters described in Table 6, which later on were processed offline after each experiment repetition. The PDR value in each node i was calculated from its P_i^{vec} vector. Note that, in each offline experiment repetition, only the `sender` and `receiver` modules were running in the AP and client nodes (i.e., there was no rate adaptation mechanism in place).

Table 6.- Offline evaluation parameters

Parameter	Definition
LQ_i	Link Quality measured between node i and the AP
P_i^{vec}	A vector of the user-plane packets received by node i
(x_i, y_i)	Virtual coordinates of node i in the ORBIT grid
TX_{AP}	Broadcast/multicast PHY transmission bit rate in the AP node

The offline evaluation of the AMuSe feedback selection mechanism comprised three different types of experiments. On the one hand, we used the first two types to validate the hypotheses introduced at the beginning of the section (H1-H4). On the other hand, the third experiment type was used to show how a rate adaptation algorithm would work in conjunction with AMuSe's feedback selection mechanism. In each experiment type (and repetition), we identified each ORBIT node by its (*row*, *column*) virtual coordinates in the ORBIT grid. Due to space constraints, only a representative subset of all experiment repetitions is discussed:

- **Different bit rates:** we fixed the AP multicast/broadcast transmission bit rate, denoted by TX_{AP}, to the different values allowed by Atheros 5212/5213 wireless cards operating in 802.11a mode (6, 9, 12, 18, 24, 36, 48, and 54 Mbps). The AP sent broadcast traffic during 10 seconds for each transmission bit rate value. We repeated these experiments 10 times at different times of the day without the presence of external noise.
- **Different noise levels:** we fixed the AP multicast/broadcast transmission bit rate to 12 Mbps and turned on the noise generator located in the vicinity of node (20,1). The ORBIT noise source was configured to generate Additive White Gaussian Noise (AWGN) for the entire spectrum in channel 40. Starting from a value of -80 dBm (very low noise power), we progressively increased

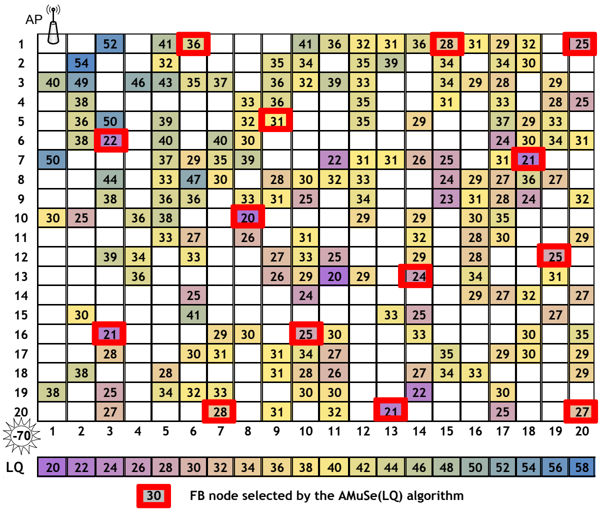
the noise power in steps of 5 dBm up to -5 dBm (high noise power). Note that the maximum noise power for the ORBIT noise generators is 0 dBm.

- **With bit rate variation:** we ran several experiment repetitions for feedback node selection with different transmission bit rate values in the access point in order to illustrate the practicality of our design and to evaluate the overall system performance.

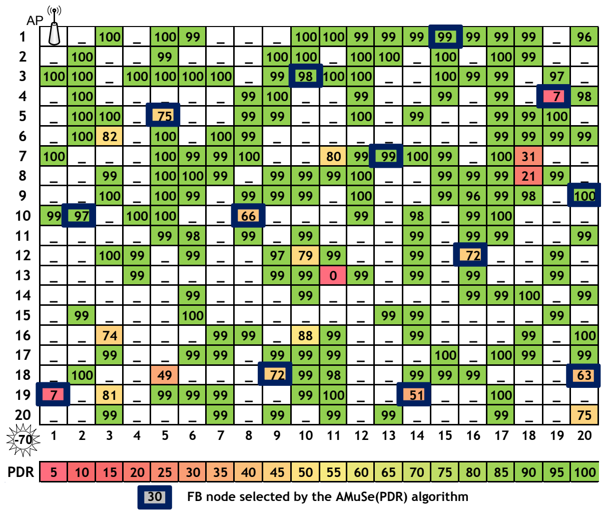
Figure 51 shows three sample heatmaps of an offline experiment repetition with TX_{AP} = 12 Mbps and various external noise powers in the vicinity of node (20,1) (further experiment repetitions showed similar results). Each heatmap shows the active nodes in the experiment, as well as their experienced LQ/PDR values. In addition, the FB nodes selected by the AP (with a D-adjacency parameter (D) of 6 meters) are also shown in the figure. Nodes marked with thick red or blue lines are FB nodes selected by the AMuSe feedback selection mechanism. Nodes with PDR = 0 are active nodes that, although reporting meaningful LQ values periodically, were unable to decode packets in the experiment repetition. For example, in Figure 51(a) and Figure 51(b), when the noise power is set at -70 dBm, node (13,11) (with PDR = 0 and LQ = 20) is such a node. These nodes were excluded from the FB node selection algorithm.

An interesting observation is that a feedback node v may have a higher PDR (or LQ) than an adjacent non-FB node u. Such a situation results from the independent-set property of the selected FB nodes and it may occur if u is D-adjacent to another FB node with lower PDR (or LQ). For example, in Figure 51(b), node (7,13) with PDR = 99% was selected as FB node although it has a neighbour (node (7,11)) with PDR = 80%. The reason is that node (7,11) is 6-adjacent to FB node (10,8) with PDR = 66%.

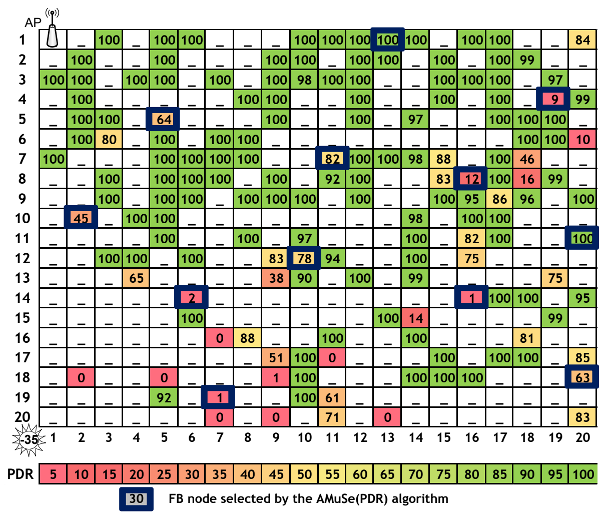
Below we proceed to evaluate hypotheses H1-H4 based on the information collected from the first two types of offline experiments described above:



(a) Link Quality heatmap, noise power = -70 dBm, D = 6 m

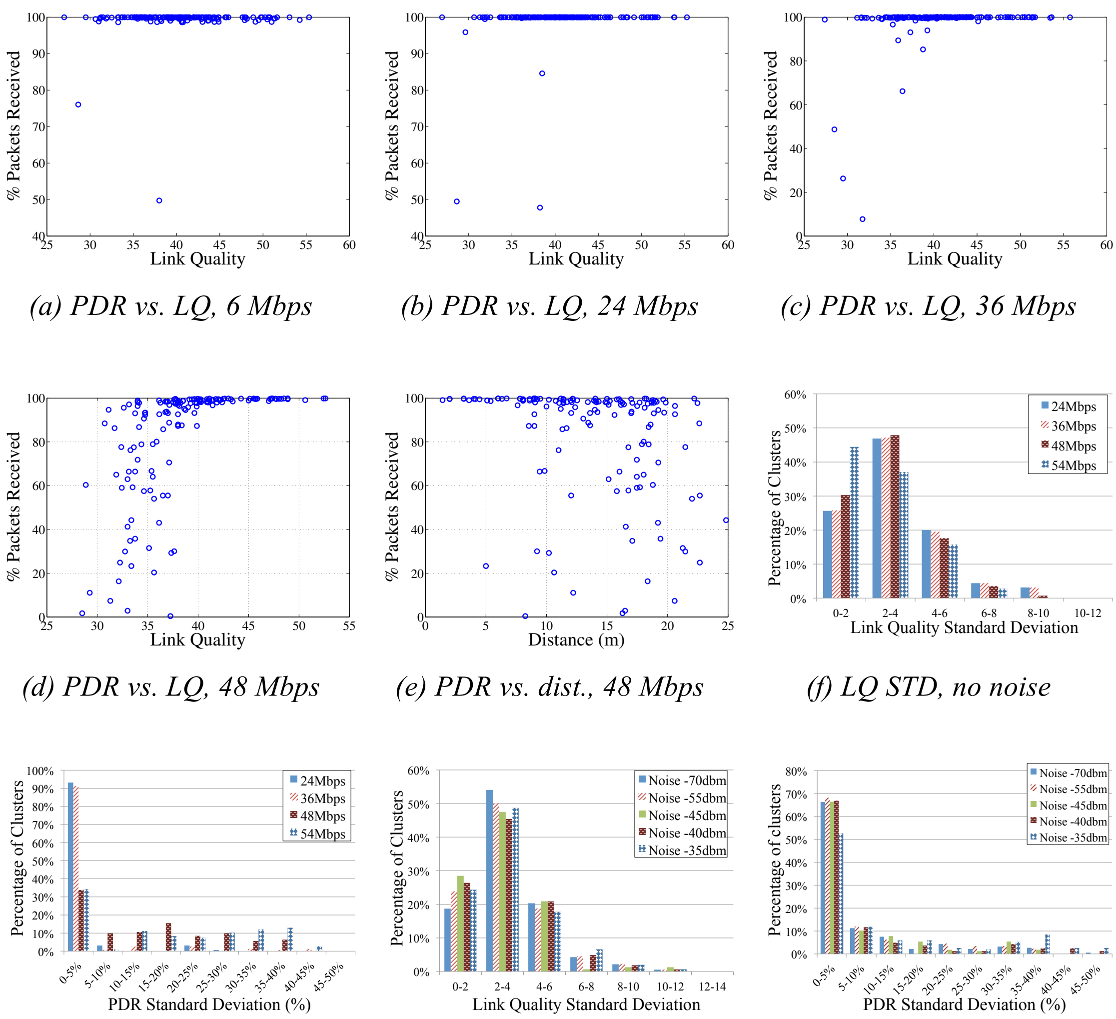


(b) PDR heatmap, noise power = -70 dBm, D = 6 m



(c) PDR heatmap, noise power = -35 dBm, D = 6 m

Figure 51.- LQ and PDR heatmaps for offline FB selection evaluation



(a) PDR vs. LQ, 6 Mbps *(b) PDR vs. LQ, 24 Mbps* *(c) PDR vs. LQ, 36 Mbps*

(d) PDR vs. LQ, 48 Mbps *(e) PDR vs. dist., 48 Mbps* *(f) LQ STD, no noise*

(g) PDR STD, no noise *(h) LQ STD, noise, 12 Mbps* *(i) PDR STD, noise, 12 Mbps*

Figure 52.- Experimental results for the offline evaluation of hypotheses H1-H3

H1 (correlation between PDR and LQ): Figure 52(a)-(d) shows the correlation between PDR and LQ in a generic wireless node for different AP transmission bit rates in the absence of external noise. Likewise, Figure 52(e) shows the correlation between a node's PDR with respect to its distance from the AP with TX_{AP} = 48 Mbps. As shown in the figures, PDR values are close to 100% for almost all nodes for bit rates up to 24 Mbps (Figure 52(a)-(b)). Some PDR degradation, though, is observed for bit rates of 36 Mbps (Figure 52 (c)). Finally, a higher variance of PDR values is seen for TX_{AP} = 48 Mbps (Figure 52(d)) and above.

As shown in Figure 52(d), the correlation between PDR and LQ is not very strong. This suggests that nodes with the same LQ values may experience significantly different PDRs. Similarly, Figure 52(e) illustrates a weak correlation between the

PDR of a node and its proximity to the AP (with TX_{AP} = 48 Mbps). In addition, some of the nodes adjacent to the AP suffer from low PDR. For instance, Figure 52(e) shows that one of the nodes 5 meters away from the AP suffers from PDR = 25%. The observed variation of PDR with LQ, as well as the variation of PDR with the distance to the AP, is consistent with prior work [112], [113], [114].

H2 (clustered nodes experience similar LQ and PDR): we calculated the standard deviation (STD) of the LQ and PDR measurements for each cluster of radius of 1.5 meters in the ORBIT grid, where each cluster contains a FB node as well as all its neighbours (between 1 and 8 nodes). Histograms of the distribution of the LQ and PDR STD for different clusters are shown in Figure 52(f) and Figure 52(g), respectively. We measured the same distributions in the presence of various noise levels and plotted the results shown in Figure 52(h) and Figure 52(i), respectively. We expect the STD across clusters to be a good measurement of how similar the PDR and the LQ values are in the AMuSe system. By comparing Figure 52(f) and Figure 52(h) we see that the LQ STD is very similar across all transmission bit rates regardless of the noise levels, as expected. Nearly 75% of the clusters have LQ STD between 0-4 dB, which confirms our assumption that adjacent nodes experience similar LQ (and RSS), and that the LQ metric does not capture the level of interference in the system.

We now consider the distribution of the PDR STD values. Figure 52(g) shows that when $TX_{AP} \leq 36$ Mbps only very few clusters show significant deviations (>5%) in PDR, as most nodes have PDR values above 99%. However, we see a significant PDR variability at higher bit rates. As shown in Figure 52(i), the presence of external noise introduces noticeable deviations (>5%) in PDR across nearly 1/3 of the clusters. To understand this we must revisit the heatmaps in Figure 51(c). It is clear that the PDR values are decreasing for nodes near the bottom-left corner of the ORBIT testbed (where the noise generator is located). The nodes that report PDR = 0 are the ones that are unable to either decode the AP beacons or, alternatively, receive multicast user-plane traffic. Disconnected nodes are not shown in the heatmaps (and, thus, not included in the variance calculations). Instead, nodes that report PDR = 0 are highlighted in red with a zero value. This becomes very noticeable as the external noise power level increases and explains the high PDR variance in Figure 52(i).

H3 (clustered nodes suffer from similar interference): Figure 51 and Figure 52(i) show that the presence of external noise has only local effects, which substantiates the need for a well-distributed set of FB nodes that report on the interference experienced by wireless receivers. Furthermore, our experiments show that increasing TX_{AP} has an impact on all client nodes. In particular, as shown in Figure 52(d) and Figure 52(e), beyond a certain transmission bit rate the PDR of many client nodes drops below 90%. This indicates the impact of the transmission bit rate (TX_{AP}) on the efficiency of the AMuSe system.

As previously explained, abnormal nodes can have a significant impact on the performance of the AMuSe system. We define an abnormal node as a client node such that its PDR is below the *abnormality threshold*, H = 90%. This is shown in Figure 52(a)-(d) as a function of TX_{AP} and LQ. As shown in the figure, PDR varies as a function of LQ for each node in a single experiment repetition with TX_{AP}={6, 24, 36, 48} Mbps, respectively (further experiment repetitions showed similar results). In general, we conclude that the number of abnormal nodes increases as TX_{AP} increases. This is consistent with the use of less robust modulation and coding schemes as the transmission bit rate values in the AP increase. In Figure 52(a)-(c), PDR values are close to 100% for the vast majority of wireless nodes for bit rates up to 36 Mbps. However, Figure 52(a) shows that, even in the extreme case of a very low TX_{AP} value (6 Mbps) and in the absence of external noise, some nodes (2 in the figure) are abnormal and suffer from low PDR.

In general, the set of abnormal nodes remains small when we increase TX_{AP} to higher bit rates (up to 36 Mbps), as shown in Figure 52(b) and Figure 52(c). However, the number of abnormal nodes increases significantly once TX_{AP} reaches 48 Mbps. Furthermore, the set of abnormal nodes is not the same across all experiment repetitions.

After discussing the impact of abnormal nodes on the AMuSe system, we now focus on comparing AMuSe's feedback selection algorithm with some other FB selection heuristics, as well as on validating hypothesis H4. In order to do so, we consider the following feedback selection algorithms:

(a) **AMuSe(LQ):** the AMuSe(LQ) algorithm is the AMuSe feedback selection mechanism described in Section 7.2.1.3, based on the LQ metric.

(b) **AMuSe(PDR):** the AMuSe(PDR) algorithm is the AMuSe feedback selection mechanism described in Section 7.2.1.3, based on the PDR metric.

(c) **K-random:** in the K-random feedback selection mechanism, the AP randomly selects a given number of nodes (k) as FB nodes.

(d) **K-best:** in the K-best feedback selection mechanism, the AP selects k nodes with the lowest LQ values as FB nodes.

As described above, each non-FB node i is associated with a FB node, denoted as FB_i. In the AMuSe system, the association takes place during cluster creation. In other feedback selection schemes, non-FB nodes are associated with the nearest FB node.

As far as the evaluation metrics are concerned, we define the *false positive ratio* of node i (denoted by FPR_i) as the fraction of multicast packets not received by non-FB node i (but reported as received correctly by its associated FB node), divided by all the transmitted packets. A high false-positive ratio implies that feedback reported by FB nodes is not accurate. This is due to high levels of lost (or incorrectly decoded) packets in non-FB nodes even if their corresponding FB node indicates otherwise.

A FB node is denoted as *satisfied* if its PDR $\geq H$ (for example, H = 90% in our experiments), since such FB nodes report a satisfying multicast/broadcast service (the *satisfaction threshold* can be properly adjusted according to the nature of the multicast stream). Instead, a non-FB node i is termed as *poorly represented* if it has a high false-positive ratio and is associated with a satisfied FB node. We characterise such a node i by $PDR_{FB_i} \geq 90\%$ and $FPR_i > 10\%$. In the ideal case, we would like FPR_i to be very low for all nodes, regardless of the state of their FB node.

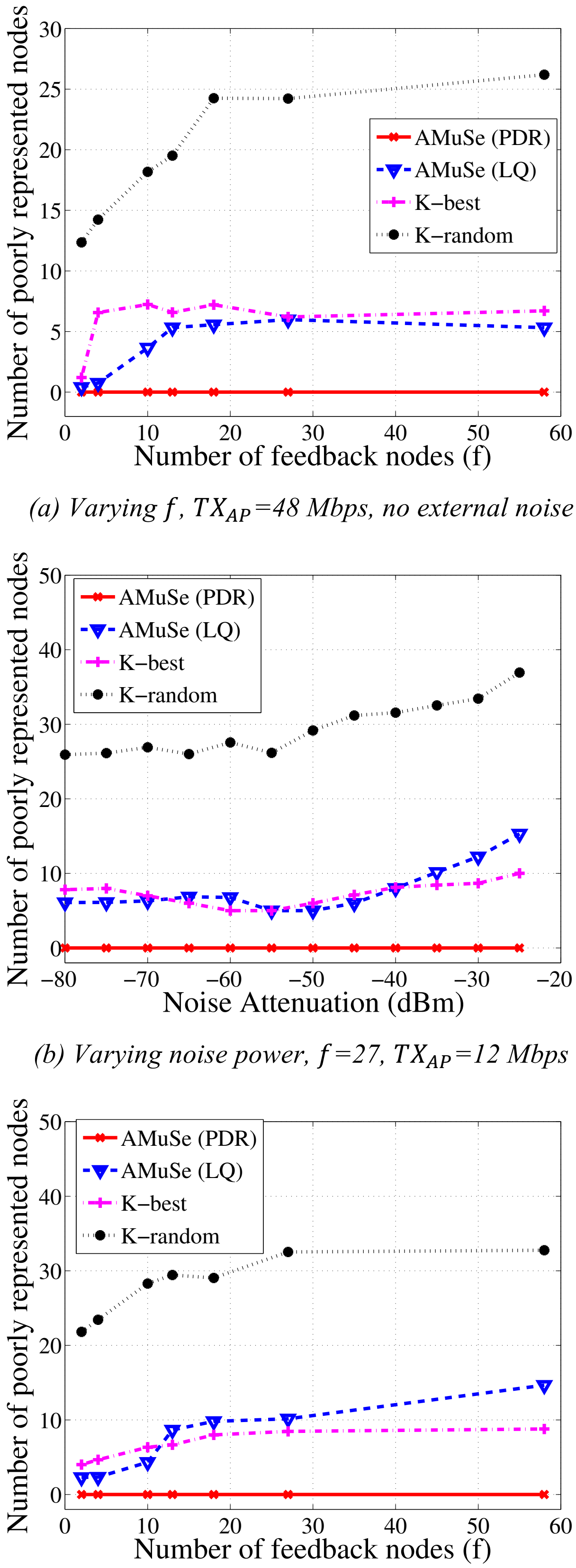
In order to provide a fair comparison across all feedback selection mechanisms, we evaluated the number of poorly represented nodes for all schemes with the same number of FB nodes (we denote this as f). Furthermore, we varied the number of selected FB nodes by changing the D-adjacency value (the cluster radius) for the

AMuSe(LQ) and AMuSe(PDR) schemes, and by changing the value of k to the corresponding values for the K-best and K-random schemes. Finally, we evaluated the sensitivity of AMuSe to various values of D. Figure 53 shows the average number of poorly represented nodes in each one of the feedback selection schemes for different values of f, TX_{AP}, and in the presence (or absence) of external noise in the ORBIT grid.

Figure 53(a) shows the number of poorly represented nodes for all feedback selection mechanisms as a function of f in the absence of external noise with TX_{AP} = 48 Mbps. Results for all feedback selection schemes have been averaged over five experiment repetitions (in the case of the K-random scheme, we have considered 20 random feedback selections to ensure a confidence level of 95%). As shown in the figure, we see that the AMuSe(PDR) mechanism does not have any poorly represented nodes for all values of f. On the other hand, AMuSe(LQ) is slightly better than the K-best feedback selection algorithm, whereas all three schemes outperform K-random.

Figure 53(b) shows the results obtained after repeating the experiments described above with external noise injection at varying levels for all feedback selection mechanisms. In this case, the number of feedback nodes (f) is fixed at 27, which is the typical number of FB nodes obtained by the AMuSe(PDR) scheme for D = 3 m. Once again, AMuSe(PDR) outperforms all other feedback selection mechanisms. On the other hand, AMuSe(LQ) performs poorly at high noise levels, which is intuitive as the LQ metric does not correlate well with PDR in the high-noise region. Similarly, in Figure 53(c) we illustrate the effects of increasing the number of feedback nodes (f) for an external noise level of -35 dBm[12]. In this scenario we see that AMuSe(PDR) outperforms all other schemes.

[12] For external noise levels higher than -35 dBm a significant number of wireless nodes stop receiving packets from the AP. This loss of synchronisation between the AP and the client nodes is caused by a high number of lost AP beacons.



(a) Varying f, TX_{AP}=48 Mbps, no external noise

(b) Varying noise power, f=27, TX_{AP}=12 Mbps

(c) Varying f, TX_{AP}=12 Mbps, noise power = -35 dBm

Figure 53.- Average number of poorly represented nodes in various scenarios

It must be noted that in the case of $f = \{2,4\}$ (two and four FB nodes, respectively) AMuSe(PDR), AMuSe(LQ), and K-best yield zero poorly represented nodes. This is because the poorest performing nodes in the network are being selected such that PDR < 90%. By definition, these are abnormal nodes and, consequently, there are no satisfied FB nodes in the network. Furthermore, this indicates that selecting too few FB nodes is not a good strategy, as the rate adaptation mechanism in the AP will eventually be dictated by a set of poorly performing abnormal nodes. For instance, Figure 51(b)-(c) shows that selecting only two or four FB nodes in the network yields PDR values lower than 30%. Clearly, these FB nodes do not provide an accurate estimation of the overall multicast QoS experienced by most of the wireless nodes. In summary, apart from not suffering from poorly represented nodes, the AMuSe(PDR) scheme has a cumulative false positive ratio of less than 1% summed over all nodes and across all experiments runs. These results validate the suitability of AMuSe's feedback selection algorithm.

Finally, we focus on the issue of location error sensitivity in feedback selection mechanisms. In order to do so, we evaluate the sensitivity of AMuSe's FB node selection algorithm to errors in node location estimation by deliberately injecting variations in the reported locations of the ORBIT nodes. These variations are picked randomly from a Gaussian distribution with $\mu = 0$, $\sigma = 7$ m, and with the D-adjacency parameter set to $D = 4.5$ m. Despite node location variations, there is no increase in the number of poorly represented nodes when AMuSe(PDR) uses the incorrect coordinates. We start observing a mild increase in the number of such nodes for $\sigma \geq 10$ m, and this levels off for higher location errors. In fact, increasing the inaccuracy in node location estimation results in AMuSe(PDR) degenerating to the K-best scheme. Obviously, this leads to low false positive ratios, but the selected FB nodes may not yield an accurate representation of the wireless network. In summary, these results confirm the suitability of FB node selection algorithms based on PDR metrics and show that AMuSe(PDR) is resilient to location error to an acceptable degree.

7.2.2.3 Online Experiments

So far we have evaluated the performance of the AMuSe feedback selection mechanism by means of offline analysis. In this section we describe the online experimental campaign carried out to evaluate the performance of a full-fledged implementation of the AMuSe system (i.e., feedback selection plus rate adaptation) in the ORBIT testbed. Prior to discussing the experimental results we describe some of the technical challenges faced during the online evaluation of the AMuSe system:

- **PDR variability:** as expected in any wireless environment, PDR estimates in client nodes are time varying, even for a fixed AP multicast transmission bit rate. Such PDR variability has an impact on the stability of the feedback list. We address this issue by defining a QoS hysteresis margin (QoS_{HYST} = 3%) for PDR measurements at client nodes in order to prevent them from volunteering as FB nodes as a result of very small variations in PDR measurements. Essentially, QoS_{HYST} provides stability to the feedback list, thus preventing FB node reconfiguration throughout the entire network.

- **Abnormal nodes:** as discussed in the previous section, even for low TX_{AP} values (and in the absence of external noise), abnormal nodes suffer from low PDR. For practical purposes, we are interested in excluding the small minority of nodes with abnormal QoS from the multicast distribution service in order to improve the overall network utilisation whilst effectively serving the majority of the wireless nodes. Thus, abnormal nodes must not represent other nodes (except themselves) and they must be easily identifiable by the AP. In order to do so, abnormal nodes can automatically set their D-adjacency parameter to 0. By doing so, neighbouring nodes will no longer consider the abnormal node as a valid FB node. In turn, this will cause the neighbouring nodes to start sending FBN-JOIN messages to become FB nodes[13]. In summary, setting the D parameter to zero guarantees that all abnormal nodes (PDR < H - QoS_{HYST}) will volunteer to become FB nodes, and that the AP will keep track of the abnormal node set, denoted by A.

- **Nodes disconnecting from the AP:** high values of TX_{AP} (in conjunction with overall system interference) can lead to some wireless nodes losing synchronisation and, thus, disconnecting from the AP. This is shown by zero values in the PDR heatmaps in Figure 51(b)-(c). High AP transmission bit

[13] Setting the D-adjacency parameter to 0 may potentially lead to FBN-JOIN message bursts in the feedback selection process. However, our experimental results show that this mechanism works well in practice without high message overhead.

rates imply higher spectrum efficiency and, thus, less robust modulation and coding schemes in the AP PHY layer (e.g., 16-QAM, 64-QAM). Analogously, low AP transmission bit rates imply lower spectrum efficiency, but more robust modulations, such as BPSK and QPSK. These changes in the AP PHY layer have an impact on the transmission bit rate of 802.11a beacon messages, which, in turn, results in some nodes losing association with the AP due to poor signal quality. In order to solve this issue, we developed a custom version of the ath5k wireless driver that allowed variable transmission bit rate selection for multicast and unicast user-plane traffic in conjunction with fixed transmission bit rate for AP beacons (6 Mbps).

With regards to the rate adaptation algorithm, we defined a simple mechanism to illustrate the positive impact of effective feedback selection and QoS measurement collection on system throughput. During the course of the online experimental campaign, we set the majority threshold X to 95%. In sum, this means that the rate adaptation algorithm collects feedback data in the AP and attempts to update the transmission bit rate periodically to ensure that *at least* X = 95% of the nodes experience a PDR above H = 90% (in fact, AMuSe tolerates a small PDR variability of QoS_{HYST} = 3%). Nodes with PDR $\geq H$ are considered as normal nodes, whereas nodes with PDR < H - QoS_{HYST} are flagged as abnormal with their D-adjacency value set to 0. On the other hand, nodes with PDR values between H - QoS_{HYST} and H keep their previous state.

By using a simple rate adaptation algorithm, the AMuSe system starts transmitting multicast/broadcast user-plane data with the lowest TX_{AP} value in the 802.11a standard (6 Mbps) and iteratively increases (or decreases) TX_{AP} according to the PDR reports received from normal and abnormal FB nodes. At each algorithm iteration (in our online experiments, once per second) AMuSe performs the following operations:

1. The rate adaptation algorithm in the AP calculates the total number of abnormal nodes in the network, $|A|$.
2. If $|A|$ is greater than 5% of the total receiver population, the AP reduces TX_{AP} to the next lower bit rate supported by the 802.11a standard. In practice, this

implies increasing the robustness of the AP PHY's modulation and coding scheme.

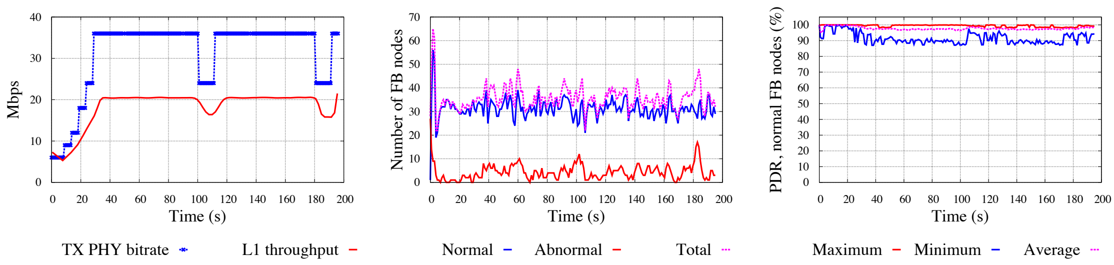
3. If $|A|$ is lower than 5%, at least X = 95% of the nodes experience high PDR above H - QoS_{HYST} = 87% and, thus, AMuSe should decide whether to increase the bit rate to the next supported value in the standard or, instead, keep the current transmission bit rate. In order to avoid undesired oscillations, TX_{AP} should only be increased if a large fraction of the client nodes ($\geq X$) will continue to benefit from high PDR ($\geq H$) with the higher bit rate. Since each increase of TX_{AP} may cause a reduction of PDR values in the receivers, we expect the above condition to hold only if more than a fraction X (= 95%) of client nodes experience very high PDR above some threshold $H_1 > X$ (e.g., in our experiments we used H_1 = 98%). Based on PDR measurements collected from FB nodes, the AP estimates the number of receivers that experience very high PDR above H_1 and, only if this exceeds X, the rate adaptation algorithm increases TX_{AP}.

As far as the online experimental setup is concerned, we used the same network topology of the offline experiments, i.e., a single multicast AP located at (1,1) and all remaining nodes in the ORBIT testbed acting as multicast/broadcast receivers. In order to deploy a full-fledged implementation of the AMuSe system, we ran concurrent instances of the `sender` and `txController` processes in the AP node, as well as the corresponding `receiver` and `feedbackManager` instances in each client node. In terms of system configuration, we set the PDR threshold H = 90% with QoS_{HYST} = 3%, and required that (at least) X = 95% of the nodes experienced PDR values higher than H. We ran multiple experiments repetitions with approximately 230 client nodes and a D-adjacency value (D) of 3 meters. Note that, in this experiment setup, the upper bound on the number of abnormal nodes ($|A|$) was $\lceil 230 \times 5\% \rceil$ = 12. We set the experiment duration to 200 seconds in order to allow the FB selection mechanism to converge to a stable set of FB nodes and, thus, show meaningful changes in the rate adaptation mechanism (as well as convergence towards higher AP transmission bit rates). In this setup we did not consider the presence of external noise sources.

In order to evaluate the performance of the AMuSe system we used the following QoS metrics:

- **PHY transmission bit rate (TX_{AP}):** this metric determines the transmission bit rate (and, hence, the associated modulation and coding scheme) used at the AP PHY layer.
- **L1 throughput at the AP:** this QoS parameter indicates the actual number of bits per second sent by the AP's physical layer during a experiment repetition. The L1 throughput ($L1_{thru}$) must not be confused with TX_{AP}. TX_{AP} determines the PHY bit rate at which the wireless card in the AP operates at a given instant (in other words, the speed at which the L1 pumps bits into the channel when the AP is working at full capacity). Instead, the L1 throughput measures the actual number of bits per second that end up being injected into the channel during normal operation. Since wireless nodes do not always operate at full capacity (e.g., they need to stop transmitting to sense the channel, wait for a back-off period after a collision, send beacons, etc.) the L1 throughput is always equal or lower than TX_{AP}.
- **Number of FB nodes:** this metric keeps track of the total number of feedback nodes in the network (both normal and abnormal) at any given instant.
- **PDR measurements:** this QoS metric determines the minimum, maximum, and average PDR measurements of the FB nodes contained in the FBN-LIST message.

Figure 54 shows the evolution of the three performance indicators over time for D = 3 meters for a given experiment repetition (as in previous cases, further repetitions show similar results). In particular, Figure 54(a) illustrates the variation of the AP transmission bit rate and its effect on L1 throughput. As shown in the figure, the system adequately converges to TX_{AP} = 36 Mbps ($L1_{thru} \approx$ 20 Mbps), a higher value than the default broadcast transmission bit rate in 802.11a (6 Mbps). Notice that some occasional drops to 24 Mbps occur at $t \approx$ 110 s and $t \approx$ 190 s. These variations in the transmission bit rate are caused by fluctuations in the PDR measurements reported by wireless nodes.



(a) TX_{AP} and L1 throughput *(b) Feedback nodes* *(c) PDR values (FB nodes)*

Figure 54.- Online experiments results, D=3

Figure 54(b) shows the number of normal and abnormal feedback nodes in the feedback list during the course of the experiment and, in turn, explains the variation of the AP transmission bit rate in Figure 54(a). Notice that the number of abnormal nodes for low TX_{AP} values (6-24 Mbps) is very small (it increases slightly and fluctuates between 0-10 nodes for $TX_{AP} \leq 36$ Mbps). Furthermore, each drop in TX_{AP} is correlated with a peak in the number of abnormal nodes. In particular, the abnormal node peaks at $t \approx 105$ s and $t \approx 185$ s ($|A| > 12$) correspond to the TX_{AP} and L1 throughput troughs at approximately the same time instants. Finally, Figure 54(c) shows how the PDR values reported by normal FB nodes (minimum, maximum, average) stay within the acceptable limits given by H - QoS_{HYST} = 87%.

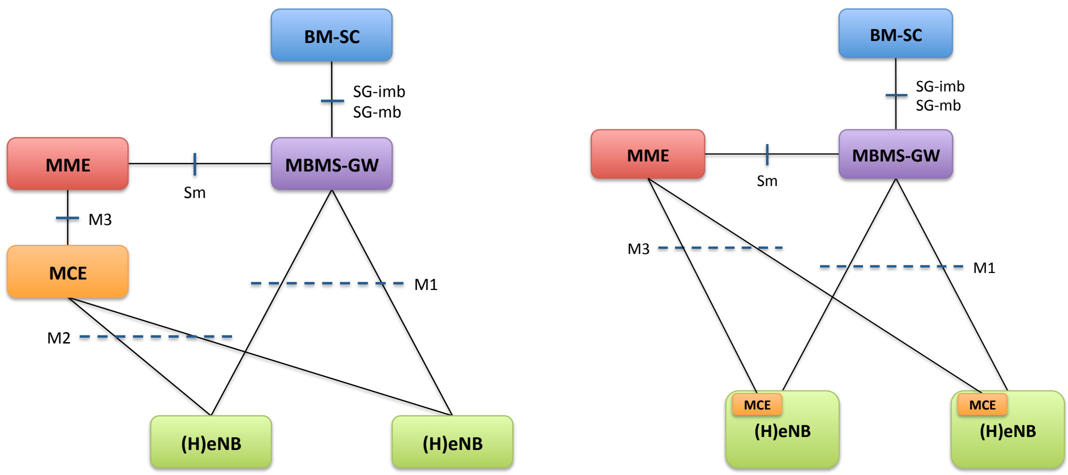
In conclusion, the live online experiments carried out in the ORBIT testbed show how the use of AMuSe's feedback selection algorithm (in conjunction with a simple rate adaptation mechanism in the AP) can lead to significant improvements in system throughput by adapting the transmission bit rate of the AP to the channel conditions and/or QoS values reported by a selected group of feedback nodes. In general, these results show the positive impact of lightweight feedback selection and rate adaptation mechanisms on wireless multicast/broadcast traffic distribution systems.

7.2.3 Implications in the MBMS Architecture

In the previous section we have described the experimental evaluation of the AMuSe system in the ORBIT tested. However, AMuSe could also be deployed in cellular systems such as LTE/LTE-A. In this section we propose an architectural and procedural framework to implement AMuSe in the 3GPP Evolved Packet System as an extension to the standard MBMS service. In particular, we focus on the implications in the network and protocol stack architecture (Sections 7.2.3.1 and 7.2.3.2), as well as in the existing MBMS network procedures (Section 7.2.3.3).

7.2.3.1 Network Architecture

The MBMS network architecture has been described in detail in Section 2.4.2.1. As previously explained, the Multi-Cell/Multicast Coordination Entity (MCE) is the EPS functional entity in charge of managing the necessary radio resources for multicast and broadcast traffic distribution. In particular, this entails configuring the appropriate modulation and coding scheme (and, thus, transmission bit rate) for the (H)eNBs participating in the MBMS service.



(a) Standalone MCE *(b) Collocated MCE(s)*

Figure 55.- MCE deployment alternatives in the MBMS network architecture

Figure 55 shows two alternatives for the deployment of the MCE functional entity in the MBMS network architecture, as defined in [7]. Each alternative has implications in the transmission bit rate/MCS used in the HeNBs participating in the MBMS session. In Figure 55(a), a single MCE has been deployed as a standalone functional entity in the MBMS network architecture. In this scenario, all HeNBs under the control of the MCE transmit multicast/broadcast user-plane data with the same MCS and bit rate configuration. Instead, Figure 55(b) depicts a network architecture where several MCE functional entities have been collocated with HeNBs in the radio access network. In this scenario, each HeNB in the MBMS service will be able to select a different {MCS, bit rate} configuration. In this section we will focus on the collocated MCE deployment alternative (Figure 55(b)) as the reference scenario for the implementation of the AMuSe system in large-scale networks of small cells.

Figure 56 illustrates the scope of AMuSe's feedback selection and rate adaptation mechanisms in the MBMS architecture. As shown in the figure, the two peer entities in the AMuSE system comprise the MCE functional entity in the HeNB and the UE. On the one hand, protocol functions related with feedback selection and rate adaptation intelligence (as those implemented by the `txController` module in Figure 50) will reside in the MCE. On the other hand, feedback-reporting functions (as those implemented by the `feedbackManager` module) will be localised in the UE. In terms of system operation, this translates into a selected group of UEs sending periodic QoS reports to the MCE functional entities in their corresponding serving HeNBs. In response, MCEs adjust the modulation and coding scheme (and, consequently, the transmission bit rate) in each HeNB. Note that HeNBs are also responsible for providing periodic information about feedback UEs in their coverage area (i.e., the feedback list). This is discussed in Section 7.2.3.3.

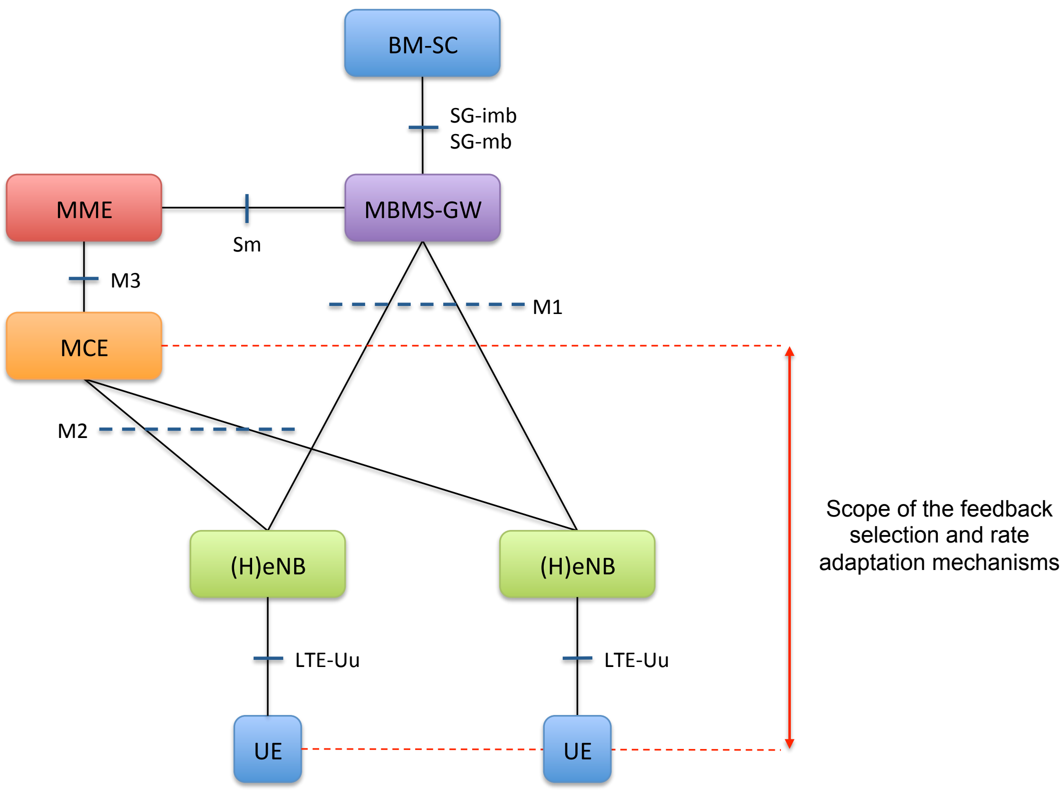


Figure 56.- Scope of AMuSe's feedback selection and rate adaptation mechanisms

7.2.3.2 Protocol Stack Architecture

In this section we propose an extension to the EPS protocol stack architecture aimed at supporting the integration of AMuSe's feedback selection and rate adaptation mechanisms in the MBMS service. In order to mitigate the impact on 3GPP standards, our proposal encapsulates AMuSe's system functions in application-level protocol entities located in the UE and the MCE[14] control-plane stacks, respectively. This is shown in Figure 57.

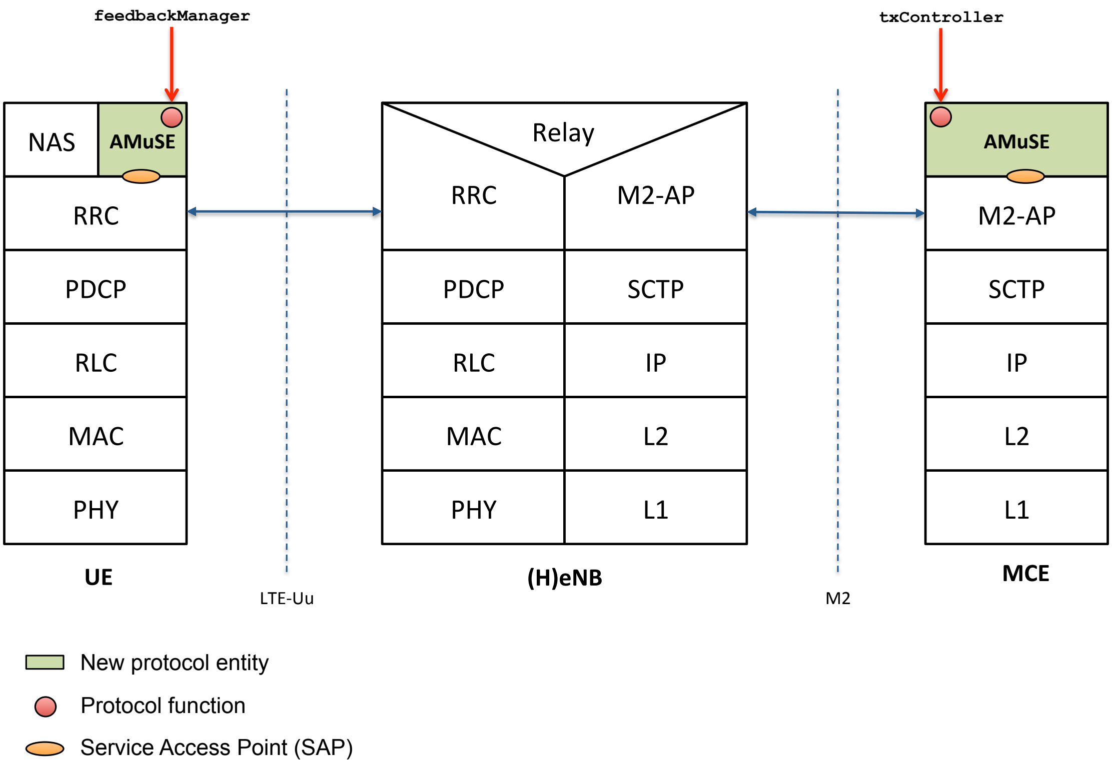


Figure 57.- Integration of AMuSe's functions in the MBMS protocol architecture

As seen in the figure, two application-layer protocol entities have been added to the standard LTE-Uu and M2 control-plane protocol stacks. These entities feature the feedback selection and rate adaptation mechanisms implemented in the `feedbackManager` and `txController` modules. In particular, the `feedbackManager` functions are localised in the UE protocol entity (left), whilst the `txController` functions reside in the MCE protocol entity (right). From an operational point of view, application-layer entities communicate with the underlying RRC and M2-AP

[14] Note that the implementation of AMuSe in the Evolved Packet System has no implications in the MBMS user-plane protocol stack.

protocols through their respective Service Access Points (SAPs). This way, the information contained in the FBN-JOIN, REPORT, and FBN-LIST protocol messages is transparently encapsulated into RRC and M2-AP messages exchanged between the UE and the HeNB, and the HeNB and the MCE, respectively. The associated network procedures are described in detail in the next section.

7.2.3.3 Network Procedures

The AMuSe system is built upon four basic network procedures, namely *FB node volunteer request*, *FB node QoS report*, *FB list announcement*, and *rate adaptation*. The first three procedures implement the protocol messages of the AMuSe system (FBN-JOIN, REPORT, and FBN-LIST, respectively). On the other hand, the *rate adaptation* procedure addresses the issue of adjusting the MCS configuration of multicast/broadcast data in the HeNB. In this section we propose a 3GPP procedural framework for the implementation of AMuSe's network procedures in the Evolved Packet System.

As shown in Figure 57, AMuSe's client and server protocol entities reside in the UE and MCE, respectively. On the client side, UEs participating in an MBMS session compute local estimations of their experienced QoS conditions (e.g., LQ, PDR, etc.) on a periodic basis. Simultaneously, on the server side, HeNB-collocated MCEs feed these measurements into AMuSe's feedback selection and rate adaptation algorithms. In this scenario, MCEs and UEs communicate with each other by using the RRC and M2-AP protocols in their hosting/serving HeNBs, respectively. Note that UEs acting as feedback nodes will operate in RRC active mode.

Figure 58 shows the message sequence chart of a generic feedback selection and rate adaptation scenario in MBMS. Standard (and newly introduced) RRC and M2-AP protocol messages have been identified in the figure for each AMuSe network procedure (*FB node volunteer request*, *FB node QoS report*, *FB list announcement*, and *rate adaptation*). As shown in the figure, the *FB node volunteer request* and *FB node QoS report* procedures are implemented by means of the standard *RRC UL Information Transfer* [15] and the newly defined *M2-AP UL Information Transfer* protocol messages. Essentially, these messages encapsulate the information contained in the AMuSe FBN-JOIN and REPORT messages, as described in Section 7.2.1.3.

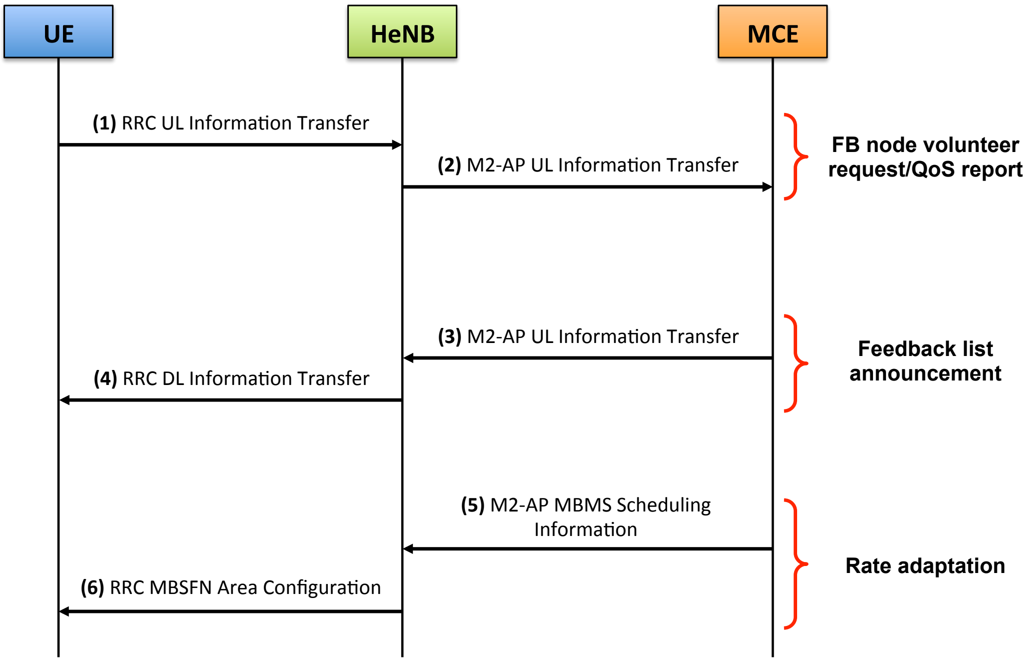


Figure 58.- Message sequence chart of a generic AMuSe scenario in MBMS

The ASN.1 structure of the *M2-AP UL Information Transfer* message is as follows:

```
ULInformationTransfer ::= SEQUENCE {
   messageType                      MessageTypeIE,
   id-MCE-MBMS-M2AP-ID              INTEGER (0..16777215),      -- 24 bits
   id-ENB-MBMS-M2AP-ID              INTEGER (0..65535),         -- 16 bits
   tmgi                             TMGI,
   ulInformationTransfer-r8         ULInformationTransfer-r8-IEs}

MessageTypeIE ::= SEQUENCE {
   procedureCode                    INTEGER (0..255),
   typeOfMessage                    ENUMERATED {
                                       Initiating Message,
                                       Successful Outcome,
                                       Unsuccessful Outcome}}

TMGI ::= SEQUENCE {
   pLMNidentity                     OCTET STRING (SIZE(3)),
   serviceID                        OCTET STRING (SIZE(3))}

ULInformationTransfer-r8-IEs ::=     SEQUENCE {
   dedicatedInfoType                CHOICE {
      dedicatedInfoAMuSe               DedicatedInfoAMuSe}}

DedicatedInfoAMuSe ::= OCTET STRING                             -- TS 36.331
```

Figure 59.- ASN.1 structure of the M2-AP UL Information Transfer message

The ASN.1 structure of the *M2-AP UL Information Transfer message* has been inherited from the *M2-AP Error Indication* and the *RRC UL Information Transfer* messages, as defined in [27] and [15], respectively. In particular, this message contains the following information elements (IE):

- **Message type:** this IE uniquely identifies the M2-AP message exchanged between the MCE and (H)eNB.
- **MCE MBMS M2AP ID:** this IE determines the MBMS service-associated logical M2 connection within an MCE functional entity.
- **(H)eNB MBMS M2AP ID:** this IE determines the MBMS service-associated logical M2 connection within a (H)eNB functional entity.
- **Temporary Mobile Group Identity (TMGI):** this IE uniquely identifies an MBMS bearer service (session).
- **Dedicated AMuSe Information:** this IE contains the corresponding protocol message in the AMuSe system, i.e., FBN-JOIN, FBN-LIST, and REPORT. In order to mitigate the impact on current 3GPP standards, the structure of the *DedicatedInfoAMuSe* IE (octet string) is identical to that of a generic payload in the standard *RRC UL Information Transfer* message.

As shown in Figure 58, the *feedback list announcement* procedure comprises two communication steps (labelled as steps 3 and 4 in the figure). First, the contents of the feedback list are sent from the MCE to the HeNB in a newly defined *M2-AP DL Information Transfer* message (note this is the downlink counterpart of the *M2-AP UL Information Transfer* message). Secondly, the HeNB encapsulates the feedback list in a standard *RRC DL Information Transfer* message and sends it over the Physical Multicast Channel (PMCH) to all UEs in the MBMS session. Finally, the RRC protocol in the UE passes the contents of the feedback list to the AMuSe protocol entity in the application layer via the corresponding SAP (see Figure 57). Figure 60 shows the ASN.1 definition of the *M2-AP DL Information Transfer* message. Note that the structure of this message (as well as the information elements contained in it) is nearly identical to that of the *M2-AP UL Information Transfer* message described in Figure 59.

```
DLInformationTransfer ::= SEQUENCE {
   messageType                    MessageTypeIE,
   id-MCE-MBMS-M2AP-ID            INTEGER (0..16777215),       -- 24 bits
   id-ENB-MBMS-M2AP-ID            INTEGER (0..65535),          -- 16 bits
   tmgi                           TMGI,
   dlInformationTransfer-r8       DLInformationTransfer-r8-IEs}

MessageTypeIE ::= SEQUENCE {
   procedureCode                  INTEGER (0..255),
   typeOfMessage                  ENUMERATED {
                                     Initiating Message,
                                     Successful Outcome,
                                     Unsuccessful Outcome}}

TMGI ::= SEQUENCE {
   pLMNidentity                   OCTET STRING (SIZE(3)),
   serviceID                      OCTET STRING (SIZE(3))}

DLInformationTransfer-r8-IEs ::=   SEQUENCE {
   dedicatedInfoType              CHOICE {
      dedicatedInfoAMuSe             DedicatedInfoAMuSe}}

DedicatedInfoAMuSe ::= OCTET STRING                            -- TS 36.331
```

Figure 60.- ASN.1 definition of the M2-AP DL Information Transfer message

Finally, we focus on the *rate adaptation* network procedure in the AMuSe system. The purpose of this procedure is to update the MCS configuration at both the HeNB and the UE in order to adjust the transmission bit rate of the multicast session. Note that both the HeNB and the UE must keep the same MCS configuration at all times in order to ensure a proper reception of the multicast/broadcast service in the UE.

As shown in Figure 58, AMuSe's *rate adaptation* procedure comprises two communication steps (labelled as 5 and 6 in the figure). First, the MCE sends a standard *M2-AP MBMS Scheduling Information* message to the HeNB. Amongst other configuration parameters, this message contains the *dataMCS* information element, i.e., the MCS settings that must be applied to the corresponding MBMS session in the HeNB. Upon reception of this message, the serving HeNB reconfigures the MCS settings for the corresponding multicast/broadcast stream and sends a standard *RRC MBSFN Area Configuration* message to the UE. As defined in [15], this message also contains the *dataMCS* IE in order to inform the UE about the new MCS settings for the corresponding MBMS session.

7.3 Research Contributions

In the course of the design and evaluation of traffic management procedures for large-scale networks of small cells, the following contributions to the literature have been made:

- Y. Bejerano, J. Ferragut, K. Guo, V. Gupta, C. Gutterman, T. Nandagopal, G. Zussman, “Experimental Evaluation of a Scalable WiFi Multicast Scheme in the ORBIT Testbed”, in Proc. GENI Research and Educational Experiment Workshop (GREE), 19-20 Mar., 2014, Atlanta, GA (United States). Invited paper.

- Y. Bejerano, J. Ferragut, K. Guo, V. Gupta, C. Gutterman, T. Nandagopal, G. Zussman, “Scalable WiFi Multicast Services for Very Large Groups”, in Proc. 21st IEEE International Conference on Network Protocols (ICNP), 7-10 Oct., 2013, Göttingen (Germany). Acceptance rate (2013): 18.3%

- J. Núñez, J. Ferragut, J. Mangues-Bafalluy, “On Stateless Routing for an All-wireless Network of Femtocells. Implications in the 3GPP Architecture”, in Proc. 73rd IEEE Vehicular Technology Conference (VTC 2011 - Spring), Workshop on Broadband Femtocell Technologies, 15-18 May, 2011, Budapest (Hungary). Acceptance rate (2011): 51.2%

Chapter 8

Conclusions

User demand for higher mobile data rates is growing at an unprecedented rate. In this context, mobile network operators are facing the challenge of providing end users with enhanced network coverage and capacity whilst avoiding expensive upgrades to their network infrastructure. In turn, this ever-increasing demand is driving network vendors towards the adoption of innovative solutions aimed at improving network coverage and capacity at a low cost.

In the last 50 years, cell size reduction has been the key contributing factor to capacity improvements in cellular networks. This has recently gained momentum with the irruption in the market of short-range, low-power cellular base stations, i.e., small cells. Research work from both the academia and the industry has primarily focused on offloading control- and user-plane traffic from macrocells onto standalone small cells in residential/small enterprise deployments. However, the widespread use of small cells by mobile network operators opens up a vast spectrum of possibilities that spans beyond traditional offloading scenarios. In this Ph.D. Thesis dissertation we have made a case for large-scale, all-wireless networks of small cells as a natural extension to standalone small cell deployments. Networks of small cells are a flexible and cost-effective solution for improving network coverage and capacity in high user-density scenarios, such as transportation hubs, sports venues, dense urban deployments, corporate premises, shopping malls, emergency-response scenarios, etc.

Networks of small cells are a disruptive concept in 3GPP networks. For instance, this shows up in the lack of a comprehensive architectural and procedural framework capable of supporting large NoS deployments in the latest release of 3GPP Technical Specifications. To this extent, the research work presented in this Ph.D. Thesis dissertation has focused on the design and evaluation of the necessary network architecture and 3GPP procedures to support the integration of large-scale networks of small cells in the Evolved Packet System with minimal impact on current 3GPP standards. In particular, we have addressed three of the issues that hinder the deployment of large-scale, all-wireless networks of small cells in the EPS: (a) the lack of a 3GPP-compliant network architecture capable of solving the scalability problems of NoS (as well as its implications in backhaul capacity), (b) the need for location management mechanisms specifically designed for NoS scenarios, and (c) the lack of efficient traffic management schemes that exploit the possibilities of a wireless multi-hop backhaul and overcome the traditional limitations of wireless multicast/broadcast systems. A brief summary of the main results and conclusions is provided below:

In Chapter 5 we have described the BeFEMTO architecture, i.e., a 3GPP-compliant framework to support large-scale deployments of networks of small cells in the Evolved Packet System. This architecture introduces a new functional entity in the NoS called the *Local Network of Small Cells Gateway* (LNGW). Amongst other functions, the LNGW confines local mobility and traffic procedures to the NoS domain. This way, control- and user-plane data generated by local network procedures (e.g., direct traffic exchange between UEs camped on the same NoS, intra-NoS handovers, Tracking Area Updates (TAUs), paging, etc.) is kept within the scope of the NoS in order to prevent congestion in the backhaul link and, potentially, the MNO's Evolved Packet Core. Furthermore, the BeFEMTO architecture allows the mobile network operator to manage the NoS as a single aggregate entity, i.e., by hiding network complexity to the EPC whilst only exposing the features the mobile operator needs to access. Finally, the proposed architecture also specifies the new (and modified) functional entities in the Evolved Packet System, as well as the corresponding 3GPP interfaces between them. To this extent, we conclude that ***massive deployments of small cells require the confinement of control- and user-plane traffic within the NoS domain in order to solve scalability and congestion issues over the backhaul link(s) and, potentially, the Evolved Packet Core.***

In Chapter 6 we have discussed the issue of local location management in networks of small cells, i.e., the process of determining the location of a UE in the NoS upon arrival of an incoming voice or data connection from the EPC. In particular, we have proposed a solution based on a self-organising Tracking Area Update procedure (in conjunction with a distributed paging mechanism) aimed at reducing location-related signalling throughout the wireless multi-hop backhaul. On the one hand, the self-organising Tracking Area List (TAL) mechanism described in Section 6.1 adjusts the size of UE-specific tracking area lists to the mobility state and paging arrival rate of each UE. Analytical results show that the proposed mechanism generates up to a 39% less location signalling traffic per UE than the conventional (i.e., static) 3GPP TAU mechanism. On the other hand, the distributed paging mechanism described in Section 6.2 improves the efficiency of the conventional paging procedure by using a combination of unicast S1-AP paging from the P-MME to a selected HeNB in the destination tracking area, followed by a sequence of local X2-AP paging operations amongst neighbouring HeNBs. Simulation results show that the proposed scheme reduces the total number of over-the-air paging messages (including retransmissions) over the wireless multi-hop backhaul in ~70% whilst keeping the paging message payload under control. Furthermore, our mechanism also reduces the total number of wireless backhaul accesses (i.e., MAC accesses) per byte in ~80% compared to the conventional 3GPP paging scheme. This has a positive impact on end-to-end paging delay (~50% reduction) and overall network congestion. In summary, ***the use of enhanced location management mechanisms (i.e., distributed paging and self-organising Tracking Area Update) results in tangible benefits for mobile network operators in terms of signalling traffic reduction in the RAN with a minimal impact on network infrastructure, network planning, and computational complexity.***

Finally, in Chapter 7 we discussed the issue of unicast and multicast traffic management in all-wireless networks of small cells. In particular, this comprised the design of a new direct bearer service to support unicast traffic between UEs camped on the same network of small cells (i.e., the *D-bearer*), as well as the use of scalable feedback selection and rate adaptation mechanisms aimed at improving multicast/broadcast traffic distribution in 3GPP MBMS services. Thus, in Section 7.1 (*Unicast Traffic Management*) we described the architectural and procedural framework needed to support direct exchange of user-plane data between LTE peers with minimal involvement from EPC entities in both static and handover scenarios. In particular, the use of D-bearers (in conjunction with distributed routing mechanisms in the underlying transport network) contributes to a more even consumption of radio and network resources throughout the entire network of small cells. Furthermore, D-bearers help to reduce user-plane traffic load on the backhaul link towards the MNO's Evolved Packet Core. As far as multicast traffic management is concerned, Section

7.2 describes a feedback selection and rate adaptation mechanism aimed at overcoming the traditional limitations of multicast/broadcast traffic distribution in wireless systems, i.e., the use low transmission bit rates (and, thus, robust modulation and coding schemes) in order to satisfy cell-edge users needs. An extensive experimental evaluation carried out in the ORBIT testbed shows that the proposed solution achieves higher system throughput by collecting and processing PDR-based feedback from a selected subset of wireless nodes. Thus, in the case of unicast traffic management, we conclude that ***local unicast traffic in a NoS requires modifications to the EPS bearer service in order to enable a more even consumption of network resources throughout the network of small cells, and to avoid potential bottlenecks in the user-plane EPC functional entities***. As far as multicast traffic management is concerned, ***the use of lightweight feedback selection and rate adaptation mechanisms in MBMS can help to improve the overall quality of experience for multicast users at the cost of introducing a small amount of feedback data and minimal computational complexity in HeNBs.***

Traffic and mobility management are two critical features of cellular networks. However, there is a plethora of open research issues that need to be addressed in order to improve the capacity, performance, and efficiency of current cellular networks. Amongst other, these comprise software-defined radio access and core networks, advanced physical layer techniques (e.g., massive MIMO), self-organising radio resource management, heterogeneous network architectures, distributed antenna systems, energy efficiency, intelligent spectrum-sharing schemes, etc. In particular, we expect that a progressive migration towards software-defined radio access and core networks (in conjunction with large-scale deployments of heterogeneous networks and intelligent spectrum sharing schemes) will play a critical role in shaping the future of next-generation cellular networks. In the current research context, both the academia and the industry share the responsibility for proposing innovative and cost-effective solutions that can help to realise the vision of ubiquitous, scalable, and fully configurable high-capacity cellular systems. We hope that the research work presented in this Ph.D. Thesis has contributed to achieving this goal, as well as to raise awareness about some of the current issues affecting cellular networks.

References

[1] Cisco Visual Networking Index (VNI), 2013-2018 [Online]. Available: http://www.cisco.com/assets/sol/sp/vni/forecast_highlights_mobile/index.html

[2] V. Chandrasekhar, J.G. Andrews, and A. Gatherer, "Femtocell networks: a survey", IEEE Communications Magazine, vol. 46, no. 9, pp. 59-67, Sept. 2008).

[3] The Small Cell Forum [Online]. Available: http://www.smallcellforum.org

[4] Broadband Evolved FEMTO Networks Project (BeFEMTO). EC Contract number INFSO-ICT-248523 [Online]. Available: http://www.ict-befemto.eu

[5] 3GPP TS 23.401, General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access [Online]. Available: http://www.3gpp.org/DynaReport/23401.htm

[6] 3GPP TS 33.106, Technical Specification Group Services and System Aspects; 3G security; Lawful Interception requirements [Online]. Available: http://www.3gpp.org/DynaReport/33106.htm

[7] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; [Online]. Available: http://www.3gpp.org/DynaReport/36300.htm

[8] Postel, J. (ed.), "Internet Protocol", STD 5, RFC 791, September 1981 [Online]. Available: http://www.rfc-editor.org/rfc/rfc791.txt

[9] 3GPP TS 25.323, Packet Data Convergence Protocol (PDCP) specification [Online]. Available: http://www.3gpp.org/DynaReport/25323.htm

[10] 3GPP TS 36.322, Radio Link Control (RLC) protocol specification [Online]. Available: http://www.3gpp.org/DynaReport/36322.htm

[11] 3GPP TS 36.321, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification [Online]. Available: http://www.3gpp.org/DynaReport/36321.htm

[12] 3GPP TS 29.281, General Packet Radio System Tunnelling Protocol User Plane (GTPv1-U) [Online]. Available: http://www.3gpp.org/DynaReport/29281.htm

[13] Postel, J. (ed.), "User Datagram Protocol", RFC 768, August 1980 [Online]. Available: http://www.rfc-editor.org/rfc/rfc768.txt

[14] 3GPP TS 24.301, Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS) [Online]. Available: http://www.3gpp.org/DynaReport/24301.htm

[15] 3GPP TS 36.331, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification [Online]. Available: http://www.3gpp.org/DynaReport/36331.htm

[16] 3GPP TS 36.413, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP) [Online]. Available: http://www.3gpp.org/DynaReport/36413.htm

[17] Stewart, R. (ed.), "Stream Control Transmission Protocol", RFC 4960, September 2007 [Online]. Available: http://www.rfc-editor.org/refs/ref4960.txt

[18] 3GPP TS 36.423, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP) [Online]. Available: http://www.3gpp.org/DynaReport/36423.htm

[19] Small Cell Forum – Press Release, "Small Cells Outnumber Traditional Mobile Base Stations" [Online]. Available: http://www.smallcellforum.org/newsstory-small-cells-outnumber-traditional-mobile-base-stations

[20] NETGEAR, Femtocell Voice Gateway DVG834GH, datasheet [Online]. Available: http://www.netgear.com

[21] Alcatel-Lucent 9760 Small Cell Family, datasheet [Online]. Available: http://www.alcatel-lucent.com/products/9760-small-cell

[22] F. Zdarsky, A. Maeder, S. Al-Sabea, and S. Schmid, "Localization of Data and Control Plane Traffic in Enterprise Femtocell Networks", in Proc. IEEE 73rd Vehicular Technology Conference (VTC Spring 2011), 15-18 May, 2011, Budapest (Hungary).

[23] 3GPP TS 23.829, Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO) [Online]. Available: http://www.3gpp.org/DynaReport/23829.htm

[24] S. Sesia, I. Toufik, and M. Baker, "LTE: The UMTS Long Term Evolution. From Theory to Practice, 2nd edition". Chichester, United Kingdom: John Wiley & Sons Ltd., 2009.

[25] 3GPP TS 23.203, Policy and charging control architecture [Online]. Available: http://www.3gpp.org/DynaReport/23203.htm

[26] 3GPP TS 25.446, MBMS synchronisation protocol (SYNC) [Online]. Available: http://www.3gpp.org/DynaReport/25446.htm

[27] 3GPP TS 36.443, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); M2 Application Protocol (M2AP) [Online]. Available: http://www.3gpp.org/DynaReport/36443.htm

[28] 3GPP TS 36.444, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); M3 Application Protocol (M3AP) [Online]. Available: http://www.3gpp.org/DynaReport/36444.htm

[29] 3GPP TR 23.859, LIPA Mobility and SIPTO at the Local Network (LIMONET) [Online]. Available: http://www.3gpp.org/DynaReport/23859.htm

[30] China Mobile – White Paper, "C-RAN: The Road Towards Green RAN. Version 2.5, October 2011" [Online]. Available: http://labs.chinamobile.com/cran/wp-content/uploads/CRAN_white_paper_v2_5_EN.pdf

[31] Alcatel-Lucent – White Paper, "LightRadio™ Network: A New Wireless Experience, v1.0 - February 2012" [Online]. Available: http://www2.alcatel-lucent.com/techzine/lightradio-network-a-new-wireless-experience

[32] Nokia Solutions and Networks – White Paper, "Liquid Radio: Let Traffic Waves Flow Most Efficiently, v1.0 – October 2013" [Online]. Available: http://it.nsn.com/sites/default/files/document/nsn_liquid_radio_white_paper.pdf

[33] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, G. Chang, "The case for re-configurable backhaul in cloud-RAN based small cell networks", in Proc. IEEE INFOCOM, 14-19 April 2013.

[34] Ericsson – Press Release, "An antenna integrated radio unit: AIR" [Online]. Available: http://www.ericsson.com/thecompany/press/releases/2011/02/1486615

[35] Intel – White Paper, "Intel Heterogeneous Network: Solution Brief" [Online]. Available: www.intel.com/content/www/us/en/communications/communications-heterogeneous-network-brief.html

[36] Texas Instruments – White Paper, "Creating cloud base stations with TI's KeyStone multicore architecture, v1.0 – October 2011" [Online]. Available: http://www.ti.com.cn/cn/lit/wp/spry183/spry183.pdf

[37] ARM – Corporate Presentation, "ARM in Networking" [Online]. Available: http://ir.arm.com/phoenix.zhtml?c=197211&p=irol-irhome

[38] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. Di Girolamo, G. Giuliani, "RAN as a service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network", in Proc. Future Network and Mobile Summit (FutureNetworkSummit), 3-5 July, 2013, Lisbon (Portugal).

[39] C. Bernardos, A. De Domenico, J. Ortin, P. Rost, D. Wübben, "Challenges of Designing Jointly the Backhaul and Radio Access Network in a Cloud-based Mobile Network", in Proc. Future Network and Mobile Summit (FutureNetworkSummit), 3-5 July, 2013, Lisbon (Portugal).

[40] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, E.-D. Schmidt, "A Virtual SDN-Enabled LTE EPC Architecture: A Case Study for S-/P-Gateways Functions", in Proc. 2013 IEEE SDN for Future Networks and Services (SDN4FNS), 11-13 November, 2013, Trento (Italy).

[41] T.K. Forde, I. Macaluso, L.E. Doyle, "Exclusive Sharing & Virtualization of the Cellular Network", in Proc. 2011 IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 3-6 May, 2011, Aachen (Germany).

[42] H. Xie, S. Tabbane, and D. J. Goodman, "Dynamic Location Area Management and Performance Analysis", in Proc. IEEE 43rd Vehicular Technology Conference, 18-20 May, 1993, Seacaucus, NJ (United States).

[43] I.F. Akyildiz and W. Wang, "On the Estimation of User Mobility Pattern for Location Tracking in Wireless Networks", in Proc. IEEE Global Telecommunications Conference (GLOBECOM), 17-21 Nov., 2002, Taipei (Taiwan).

[44] J.H. Na, H. Cho, S. Lee, Y. Chung, and S.H. Kim, "Adaptive Location Update and Paging Scheme for Mobile Broadband Wireless Access Networks", in Proc. IEEE Wireless Communications and Networking Conference, 5-8 Apr., 2009, Budapest (Hungary).

[45] S. M. Razavi and D. Yuan, "Performance Improvement of LTE Tracking Area Design: a Re-optimization Approach," in Proc. 6th ACM International Symposium on Mobility Management and Wireless Access, 27-31 Oct., 2008, Vancouver (Canada).

[46] S. M. Razavi, D. Yuan, F. Gunnarsson, and J. Moe, "Exploiting Tracking Area List for Improving Signaling Overhead in LTE," in Proc. IEEE 71st Vehicular Technology Conference (VTC Spring 2010), 16-19 May, 2010, Taipei (Taiwan).

[47] S. M. Razavi, D. Yuan, F. Gunnarsson, and J. Moe, "Dynamic Tracking Area List Configuration and Performance Evaluation in LTE", in Proc. IEEE Globecom Workshops, 6-10 Dec., 2010, Miami, FL (United States).

[48] Y. Watanabe and M. Yabusaki, "Mobility/Traffic Adaptive Location Management", in Proc. IEEE 56th Vehicular Technology Conference, 24-28 Sept. 2002, Vancouver (Canada).

[49] P. Das and A. Chandra, "Location Management in Wireless Networks: A Survey", in Proc. IEEE World Congress on Information and Communication Technologies, 11-14 Dec., 2011, Mumbai (India).

[50] D.J. Lee and D.H. Cho, "Sequential Paging Schemes Considering Two Classes of Paging Traffic", IEEE Electronics Letters, vol.35, no.14, pp.1138-1139, Jul. 1999.

[51] A. Abutaleb and V. O. K. Li, "Paging Strategy Optimization in Personal Communication Systems", ACM/Kluwer Wireless Networks, vol. 3, no. 3, pp. 195-204, 1997.

[52] S. Mishra and O.K. Tonguz, "Analysis of Intelligent Paging in Personal Communication Systems", IEEE Electronics Letters, vol.34, no.1, pp.12-13, Jan. 1998.

[53] K. Lam, B. Y. Liang, and C. Zhang, "On Using Handoff Statistics and Velocity for Location Management in Cellular Wireless Networks", ACM/Oxford The Computer Journal, vol. 48, no. 1, pp. 84-100, Jan. 2005.

[54] Y. Xiao, H. Chen, and M. Guizani, "Performance Evaluation of Pipeline Paging Under Paging Delay Constraint for Wireless Systems", IEEE Transactions on Mobile Computing, vol. 5, no. 1, pp. 64-76, Jan. 2006.

[55] H. Zang and J. Bolot, "Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks", in Proc. 13th ACM International Conference on Mobile Computing and Networking (MOBICOM), 9-14 Sept., 2007, Montreal (Canada).

[56] E. Halepovic and C. Williamson, "Characterizing and Modeling User Mobility in a Cellular Data Network", in Proc. 2nd ACM International Workshop on

Performance Evaluation of Wireless Ad-hoc, Sensor, and Ubiquitous Networks (PE-WASUN). 13 Oct., 2005, Montreal (Canada).

[57] C. Williamson, E. Halepovic, H. Sun, Y. Wu, "Characterization of CDMA2000 Cellular Data Network Traffic", in Proc. IEEE Conference on Local Computer Networks, 17 Nov., 2005, Washington D.C. (United States).

[58] 3GPP TS 36.331, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification [Online]. Available: http://www.3gpp.org/DynaReport/36331.htm

[59] 3GPP TS 22.220, Service requirements for Home Node B and Home eNode B [Online]. Available: http://www.3gpp.org/DynaReport/22220.htm

[60] K. Samdanis, T. Taleb, S. Schmid, "Traffic Offload Enhancements for eUTRAN", IEEE Communications Surveys & Tutorials, vol.14, no.3, pp.884-896, 2012.

[61] C.B. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: A tutorial", IEEE Communications Magazine, vol.50, no.6, pp.46-53, June 2012.

[62] J.-W. Wu, C.-W. Cheng, S.-R. Yang, C.-H. Gan, Y.-B. Chen, "uLIPA: A Universal Local IP Access Solution for 3GPP Mobile Networks", in Proc. 9th International IEEE Wireless Communications and Mobile Computing Conference (IWCMC), 1-5 July, 2013, Cagliari (Italy).

[63] 3GPP TR 22.803, Feasibility Study for Proximity Services (ProSe) [Online]. Available: http://www.3gpp.org/DynaReport/22803.htm

[64] 3GPP TS 22.278, Service requirements for the Evolved Packet System (EPS) [Online]. Available: http://www.3gpp.org/DynaReport/22278.htm

[65] 3GPP TS 22.115, Service aspects; Charging and billing [Online]. Available: http://www.3gpp.org/DynaReport/22115.htm

[66] M. J. Yang, S. Y. Lim, H. J. Park, N. H. Park, "Solving the Data Overload: Device-to-Device Bearer Control Architecture for Cellular Data Offloading", IEEE Vehicular Technology Magazine, vol.8, no.1, pp.31-39, March 2013.

[67] B. Raghothaman, E. Deng, R. Pragada, G. Sternberg, T. Deng, K. Vanganuru, "Architecture and Protocols for LTE-based Device to Device Communication", in Proc. International Conference on Computing, Networking and Communications (ICNC), 28-31 Jan., 2013, San Diego (United States).

[68] K. Doppler, M. Rinne, C. Wijting, C.B. Ribeiro, K. Hugl, "Device-to-Device Communication as an Underlay to LTE-Advanced Networks", IEEE Communications Magazine, vol.47, no.12, pp.42-49, Dec. 2009.

[69] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, Z. Turányi, "Design Aspects of Network Assisted Device-to-Device Communications" IEEE Communications Magazine, vol.50, no.3, pp.170-177, March 2012.

[70] H. Ishii, X. Cheng, S. Mukherjee, B. Yu, "An LTE Offload Solution Using Small Cells with D2D Links", in Proc. IEEE International Conference on Communications (ICC), 9-13 June, 2013, Budapest (Hungary).

[71] A. Pyattaev, K. Johnsson, S. Andreev, Y. Koucheryavy, "Proximity-Based Data Offloading via Network Assisted Device-to-Device Communications", in Proc. 77th IEEE Vehicular Technology Conference (VTC Spring), 2-5 June 2013, Dresden (Germany).

[72] A. Pyattaev, K. Johnsson, S. Andreev, Y. Koucheryavy, "3GPP LTE Traffic Offloading onto WiFi Direct", in Proc. IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 7-10 April, 2013, Shanghai (China).

[73] R. Chandra, S. Karanth, T. Moscibroda, V. Navda, J. Padhye, R. Ramjee, and L. Ravindranath, “DirCast: a practical and efficient Wi-Fi multicast system”, in Proc. IEEE International Conference on Network Protocols (ICNP), 13-16 Oct., 2009, Princeton, NJ (United States).

[74] N. Choi, Y. Seok, T. Kwon, and Y. Choi, “Leader-based multicast service in IEEE 802.11v networks,” in Proc. IEEE Consumer Communications and Networking Conference (CCNC), 9-12 Jan., 2010, Las Vegas, NV (United States).

[75] J. K. Kuri and S. Kumar, “Reliable multicast in multi-access wireless LANs,” ACM/Kluwer Wireless Networking, vol. 7, no. 4, pp. 359–369, 2001.

[76] M.-T. Sun, L. Huang, A. Arora, and T.-H. Lai, “Reliable MAC layer multicast in IEEE 802.11 wireless networks,” in Proc. IEEE International Conference on Parallel Processing (ICPP), 21 Aug., 2002, Vancouver (Canada).

[77] X. Wang, L. Wang, Y. Wang, Y. Zhang, and A. Yamada, “Supporting MAC layer multicast in IEEE 802.11n: Issues and solutions,” in Proc. IEEE Wireless

Communications and Networking Conference (WCNC), 5-8 Apr. 2009, Budapest (Hungary).

[78] O. Alay, T. Korakis, Y. Wang, and S. Panwar, “Dynamic rate and FEC adaptation for video multicast in multi-rate wireless networks,” ACM/Springer Mobile Networks and Applications, vol. 15, no. 3, pp. 425–434, 2010.

[79] H.-T. Chiao, S.-Y. Chang, K.-M. Li, Y.-T. Kuo, and M.-C. Tseng, “WiFi multicast streaming using AL-FEC inside the trains of high-speed rails,” in Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 27-29 Jun., 2012, Seoul (South Korea).

[80] M. Wu, S. Makharia, H. Liu, D. Li, and S. Mathur, “IPTV multicast over wireless LAN using merged hybrid ARQ with staggered adaptive FEC,” IEEE Transactions on Broadcasting, vol. 55, no. 2, pp. 363 –374, 2009.

[81] W.-S. Lim, D.-W. Kim, and Y.-J. Suh, “Design of efficient multicast protocol for IEEE 802.11n WLANs and cross-layer optimization for scalable video streaming,” IEEE Transactions on Mobile Computing, vol. 11, no. 5, pp. 780 – 792, 2012.

[82] Z. Feng, G. Wen, C. Yin, and H. Liu, “Video stream groupcast optimization in WLAN,” in Proc. IEEE International Conference on Internet Technology and Applications (ITA), 20-22 Aug., 2010, Wuhan (China).

[83] S. Sen, N. K. Madabhushi, and S. Banerjee, “Scalable WiFi media delivery through adaptive broadcasts”, in Proc. ACM 7th USENIX Conference on Networked Systems Design and Implementation (NSDI), 28-30 Apr., 2010, San Jose, CA (United States).

[84] Y. Park, C. Jo, S. Yun, and H. Kim, “Multi-room IPTV delivery through pseudo-broadcast over IEEE 802.11 links,” in Proc. IEEE 71st Vehicular Technology Conference (VTC Spring 2010), 16-19 May, 2010, Taipei (Taiwan).

[85] Z. Li and T. Herfet, “HLBP: a hybrid leader based protocol for MAC layer multicast error control in wireless LANs,” in Proc. IEEE Global Telecommunications Conference (GLOBECOM), 30 Nov. – 4 Dec., 2008, New Orleans, LO (United States).

[86] ISO/IEC, “Information technology — Dynamic adaptive streaming over HTTP (DASH)”, ISO/IEC 23009-1:2012(E) International Standard [Online]. Available:

http://standards.iso.org/ittf/PubliclyAvailableStandards/c057623_ISO_IEC_23009-1_2012.zip

[87] 3GPP TS 26.346, “Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs” [Online]. Available: http://www.3gpp.org/DynaReport/26346.htm

[88] INFSO-ICT-248523 BeFEMTO – Deliverable D2.2, v1.0, “The BeFEMTO System Architecture” [Online]. Available: http://www.ict-befemto.eu

[89] J. Núñez, “Self-Organized Backpressure Routing for the Wireless Mesh Backhaul of Small Cells”. Ph.D. Dissertation, Computer Architecture Department, Universitat Politècnica de Catalunya – BarcelonaTech, Barcelona, Spain, 2014.

[90] INFSO-ICT-248523 BeFEMTO – Deliverable D3.2, v1.0, “Promising Interference and Radio Management Techniques for Indoor Standalone Femtocells” [Online]. Available: http://www.ict-befemto.eu

[91] INFSO-ICT-248523 BeFEMTO – Deliverable D5.2, v1.0, “Femtocells access control, networking, mobility and management mechanisms (final)” [Online]. Available: http://www.ict-befemto.eu

[92] INFSO-ICT-248523 BeFEMTO – Deliverable D4.3, v1.0, “Multi-cell RRM and self-optimisation for networked, fixed relay and mobile femtocells” [Online]. Available: http://www.ict-befemto.eu

[93] J. Núñez, J. Ferragut, J. Mangues-Bafalluy, “On Stateless Routing for an All-wireless Network of Femtocells. Implications in the 3GPP Architecture”, in Proc. 73rd IEEE Vehicular Technology Conference (VTC 2011 - Spring), Workshop on Broadband Femtocell Technologies, 15-18 May, 2011, Budapest (Hungary).

[94] 3GPP TS 36.902, “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions” [Online]. Available: http://www.3gpp.org/DynaReport/36902.htm

[95] A. Jamalipour, “The Wireless Mobile Internet. Architectures, Protocols and Services, 1st ed.”, Chichester, United Kingdom: John Wiley and Sons, Ltd., 2003.

[96] 3GPP TS 24.008, Mobile radio interface Layer 3 specification; Core network protocols; [Online]. Available: http://www.3gpp.org/DynaReport/24008.htm

[97] A. Chandra, K. Mal, "Genetic Algorithm Based Optimization for Location Update and Paging in Mobile Networks", in Proc. Asian Applied Computing Conference (AACC), 29-31 Oct., 2004, Kathmandu (Nepal).

[98] B. Liang, Z.J. Haas, "Predictive Distance-Based Mobility Management for PCS Networks", in Proc. IEEE 18th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), 21-25 Mar., 1999, New York, NY (United States).

[99] Y. Tseng, L. Chen, M. Yang, J. Wu, "A Stop-or-Move Mobility Model for PCS Networks and its Location-Tracking Strategies", ACM Computer Communications, vol. 26, no. 12, pp. 1288-1301, 2003.

[100] H. Prüfer, "Neuer Beweis eines Satzes über Permutationen". Archiv für Mathematik und Physik, vol. 27, pp. 742–744, 1918.

[101] 3GPP TS 36.413, Evolved Universal Terrestrial Radio Access Network (E-UTRAN); S1 Application Protocol (S1AP) [Online]. Available: http://www.3gpp.org/DynaReport/36413.htm

[102] The ns-3 network simulator [Online]. Available: http://www.nsnam.org

[103] Qualcomm – White Paper, "eMBMS Technology Overview" [Online]. Available: http://s3.amazonaws.com/sdieee/222-eMBMS_tech_overview_IEEE_112712.pdf

[104] The CREW Project LTE/LTE-A Testbed [Online]. Available: http://www.crew-project.eu/portal/ltelte-advanced-testbed

[105] The PerformLTE Testbed (FP7 Fed4Fire Integrating Project) [Online]. Available: http://www.fed4fire.eu/testbeds/performlte.html

[106] J.-T.J. Hung, C.-Y. Ku, H.-C. Cheng, L.-P. Tung, B.-S.P. Lin, "UE Performance Evaluation on LTE Indoor and Outdoor Testbeds", in Proc. International Symposium on Wireless and Pervasive Computing (ISWPC), 20-22 Nov., 2013, Taipei (Taiwan).

[107] The ORBIT Testbed [Online]. Available: http://www.orbit-lab.org

[108] D. Raychaudhuri, I. Seskar, M. Ott, S. Ganu, K. Ramachandran, H. Kremo, R. Siracusa, H. Liu, M. Singh, "Overview of the ORBIT Radio Grid Testbed for Evaluation of Next-Generation Wireless Network Protocols," in Proc. IEEE Wireless Communications and Networking Conference (WCNC), 13-17 Mar., 2005, New Orleans, LA (United States).

[109] D. Aguayo, J. Bicket, S. Biswas, G. Judd, R. Morris, "Link-Level Measurements From an 802.11b Mesh Network", in Proc. ACM Special Interest Group on Data

Communication (SIGCOMM), 30 Aug. - 3 Sep., 2004, Portland, OR (United States).

[110] M.V. Marathe, H. Breu, H.B. Hunt III, S.S. Ravi, D.J. Rosenkrantz, "Simple Heuristics for Unit Disk Graphs", Wiley Networks, vol. 25, no. 2, pp. 59-68, 1995.

[111] P. Mirowski, H. Steck, P. Whiting, R. Palaniappan, M. MacDonald, T.K. Ho, "KL-Divergerce Kernel Regression for Non Gaussian Fingerprint Based Localization", in Proc. 5th International Conference on Indoor Positioning and Indoor Navigation (IPIN), 21 - 23 Sep., 2011, Guimarães (Portugal).

[112] H. Rahul, F. Edalat, D. Katabi, C.G. Sodini, "Frequency-Aware Rate Adaptation and MAC Protocols", in Proc. ACM 15th International Conference on Mobile Computing and Networking (MOBICOM), 20 - 25 Sep., 2009, Beijing (China).

[113] A. Vlavianos, L. Law, I. Broustis, S. Krishnamurthy, M. Faloutsos, "Assessing Link Quality in IEEE 802.11 Wireless Networks: Which Is the Right Metric?", in Proc. IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 15 - 18 Sep., 2008, Cannes (France).

[114] S. Kaul, M. Gruteser, I. Seskar, "Creating Wireless Multi-Hop Topologies on Space-Constrained Indoor Testbeds Through Noise Injection", in Proc. IEEE/Create-Net 2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TridentCom), 1 - 3 Mar., 2006, Barcelona (Spain).