Coordinated Control and Spectrum Management for 5G Heterogeneous Radio Access Networks

Grant Agreement No. 671639

Call: H2020-ICT-2014-2

Deliverable D3.2

Final report on physical and MAC layer modelling and abstraction

Version:	1.0
Due date:	01.12.2017
Delivered date:	13.12.2017
Dissemination level:	PU

The project is co-funded by

Authors

Nidal Zarifeh, Mai Alissa (UDE) Editors-in-Chief; Antti Anttonen, Aarne Mämmelä, Tao Chen (VTT); Mariana Goldhamer (4GC); Junquan Deng, Sergio Lembo, Olav Tirkkonen (AALTO); Dimitri Marandin (CMA); Shah Nawaz Khan, Roberto Riggio (CNET); Chia-Yu Chang (EUR); George Agapiou (OTE); Pawel Kryszkiewicz, Adrian Kliks (PUT); Per Kreguer, Akhila Rao, Rebecca Steinert (SICS); Antonio Cipriano, Dorin Panaitopol (TCS); Sundar Daniel Peethala (UDE).

Coordinator

Dr. Tao Chen

VTT Technical Research Centre of Finland Ltd

Tietotie 3

02150, Espoo

Finland

Email: tao.chen@vtt.fi

Disclaimer

Acknowledgement

This report is funded under the EC H2020 5G-PPP project COHERENT, Grant Agreement No. 671639.

Version history

Version	Date	Remarks
0.0	07.03.2017	Draft ToC created
0.1	13.09.2017	Full integration of sections (2, 3.4,3.5, 3.6, 5). Missing (1, 3.1, 3.2, 3.3, 4, 6)
0.2	20.11.2017	Complete for external review
0.3	21.11.2017	Complete for external review with minor modifications in conclusion and formatting
1.0	13.12.2017	Finalized after external review

Summary

This deliverable contains the final results of the physical and medium access control layers modelling and abstraction. To avoid duplication, some contributions in D3.1 are not included here. Thus, D3.2 is not a stand-alone document, but is intended to be attached to D3.1, as they present together the technical output of this work package.

The objective of COHERENT network abstractions in WP3 is two-fold: i) to expose the network state and user conditions for network control applications via effective aggregation principles and ii) to provide means to express performance goals and configurations in a unified manner that can be automatically translated down to RAT-specific control actions. This can be mapped into six objectives in WP3, and divided into three tasks: abstraction and coordination of wireless access networks in HMN, abstraction and control of user cooperation, and network coverage extension and mobility management

To fulfil the six objectives, we have evaluated novel abstraction applications using the network graph in 16 technical contributions on coordinated control and communication in 5G. Also we identified and derived new expressions for 21 abstractions enabling resource coordination and control across heterogeneous RATs at both the RTC and C3 levels using combinations of existing and novel metrics. In a unified table we list common abstractions developed for various controller applications for resource coordination and QoS assurance, load balancing, and D2D communication. This deliverable includes the proposed abstractions which are RAT-agnostic, in the sense that control actions can be expressed at a high-level of abstraction and translated down to RAT-specific control actions. Four of the contributions were selected for joint prototyping and/or SDK code release.

Also WP3 proposed an X2-like protocol for control and coordination of 3GPP RATs and WiFi. A significant part of the contributions to 3GPP and ETSI BRAN are based on this document content, and in particular the proposed protocol, coupling directly WP3 output to WP7 efforts.

The gain of using WP3 concepts and abstractions is evaluated by the integration of a subset of the proposed methods in other WPs, especially the demonstration in WP6.

Table of contents

List of abbreviations

3GPP	Third Generation Partnership Project
5G	5th Generation
5G PPP	5th Generation Public-Private Partnership
5QI	5G QoS Indicator
ABS	Almost Blank Subframes
AC	Access Categories
ACIR	Adjacent Channel Interference Ratio
ACK	Acknowledgement
ACLR	Adjacent Channel Leakage Ratio
ACS	Adjacent Channel Selectivity
AIF	Antenna Interface
AM	Acknowledged Mode
AMD	Acknowledged Mode Data
AMF	Access and Mobility Management Function
ANC	Available Node Capacity
ANDSF	Access Network Discovery and Selection Function
AoA	Angle of Arrival
AoD	Angle of Departure
AP	Access Point
API	Application Programming Interface
ARQ	Automatic Repeat reQuest
AWGN	Additive White Gaussian Noise
BBU	Baseband Unit
BCH	Broadcast Channel
BER	Bit Error Rate
BH	BackHaul

BLER	Block Error Rate
BPSK	Binary Phase Shift Keying
BS	Base Station
BSR	Buffer Status Report
C3	Central Controller and Coordinator
CA	Carrier Aggregation
CAP	Consistency Availability Partition
CB	Coordinated Beamforming
CCA	Clear Channel Assessment
CCA-ED	Clear Channel Assessment- Energy Detection
CCTV	Closed Circuit Television
CDF	Cumulative Distribution Function
CL	Coupling Loss
CN	Cognitive Network
CoMP	Coordinated Multipoint
CP	Control Plane
CPRI	Common Public Radio Interface
CQI	Channel Quality Indicator
CQM	Channel Quality Map
CRC	Cyclic Redundancy Check
CRE	Cell Range Expansion
CRS	Cell-Specific Reference Signal
CS	Coordinated Scheduling
CSG	Closed Subscriber Group
CSI	Channel State Information
CSI-RS	Channel State Information Reference Signal

CSMA/CA	Carrier Sense Multiple Access/Collision Avoidance
CU	Central Unit
CW	Continuous Wave
CW	Contention Window
D2D	Device to Device
DA2GC	Direct Air to Ground Communications
DAS	Distributed Antenna System
DCI	Downlink Control Information
DCS	Dynamic Cell Selection
DFN	Direct Frame Number
DFT	Discrete Fourier Transform
DL	DownLink
DL-SCH	DownLink Shared CHannel
DMS	Direct Multicast Service
DRB	Data Radio Bearer
DSSS	Direct Sequence Spread Spectrum
DTC	Distributed Target Computation
DU	Distributed Unit
EARFCN	E-UTRA Absolute Radio Frequency Channel Number
EDP	Energy Depletion Probability
EMA	Exponential Moving Average
eMBB	enhanced Mobile Broadband
eNB or eNodeB	Evolved Node B
EP	Elementary Procedure
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute

E-UTRA	Evolved UMTS Terrestrial Radio Access
eV2X	Enhanced Vehicle-to-X communication
EW	East West interface
EVM	Error Vector Magnitude
EWMA	Exponentially Weighted Moving Average
FBMC	Filter Bank Multicarrier
FDD	Frequency Division Duplexing
FFT	Fast Fourier Transform
FH	Frequency Hopping
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communication
H2H	Human to Human
HARQ	Hybrid Automatic Repeat reQuest
HETNET, HetNet	Heterogeneous Network
HMN	Heterogeneous Mobile Network
HO	HandOver
ICIC	Intercell Interference Coordination
IDFT	Inverse Discrete Fourier Transform
IoT	Internet of Things
IP	Internet Protocol
ISD	InterSite Distance
ISO	International Organization for Standardization
JT	Joint Transmission
LAA	Licensed Assisted Access
LAN	Local Area Network

LBT	Listen Before Talk
LC	Logical Channel
LCG	Logical Channel Group
LE	Licence Exempt
LNV	Local Network View
LOS	Line Of Sight
LSB	Least Significant Bit
LTE	Long Term Evolution
LTE-A	LTE-Advanced
M2M	Machine to Machine
MAC	Medium Access Control
MBS	Macro Base Station
MCPTT	Mission Critical Push to Talk
MCS	Modulation and Coding Scheme
MF	Monitoring Function
MIB	Master Information Block
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MoM	Method of Moments
MR	Multicast Receptor
MSC	Message Sequence Chart
NACK	Negative Acknowledgement
NAS	Non Access Stratum
NB	NorthBound
NG	Network Graph
NGMN	Next Generation Mobile Networks

NIF	Network Information Function
NLOS	Non Line Of Sight
NOS	Network Operating System
NS	Network Slicer
NSSAI	Network Slice Selection Assistance Information
NSSP	Network Slice Selection Policy
NZP	Non-Zero Power
OAM	Operation And Maintenance
OBSAI	Open Base Station Architecture Initiative
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OOB	Out-of-Band
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDR	Packet Delivery Ratio
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
PER	Packet Error Rate
PF	Plane Function
PHY	Physical Layer
PMF	Probability Mass Function
PMR	Professional Mobile Radio
PRACH	Physical Random Access Channel
PRB	Physical Resource Block
ProSe	Proximity Services
PS	Public Safety

PSBCH	Physical Sidelink Broadcast CHannel
PSCCH	Physical Sidelink Control Channel
PSD	Power Spectral Density
PSDCH	Physical Sidelink Discovery CHannel
PSS	Primary Synchronization Signal
PSSCH	Physical Sidelink Shared Channel
PSSS	Primary Sidelink Synchronization Signal
PU	Primary User
PUSCH	Physical Uplink Shared Channel
QAM	Quadrature Amplitude Modulation
QCI	QoS Class Identifier
QoE	Quality of Experience
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RCPI	Received Channel Power Indicator
RDP	Resource Depletion Probability
RI	Rank Indicator
RLAN	Radio Local Area Network
RLC	Radio Link Control
RNTI	Radio Network Temporary Identifier
RNTP	Relative Narrowband Transmit Power
RoHC	Robust Header Compression
RP	Resource Pool

RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSSI	Received Signal Strength Indicator
RT	Radio Transceiver
RTC	Real-Time Controller
RTP	Real Transmission Point
RX	Receiver
SAE	System Architecture Evolution
SB	SouthBound
SBCH	Sidelink Broadcast Channel
SBS	Small Base Station
SC-FDMA	Single Carrier- Frequency Division Multiple Access
SCI	Sidelink Control Information
SCTP	Stream Control Trasmission Protocol
SDK	Software Development Kit
SDM	Space Division Multiplexing
SDN	Software Defined Networking
SD-RSRP	Sidelink Discovery Reference Signal Received Power
SDU	Service Data Unit
SFN	System Frame Number
SG	Serving Gateway
SIB	System Information Block
SIMO	Single Input Multiple Output

SINR	Signal to Interference plus Noise Ratio
SIR	Signal to Interference Ratio
SISO	Single Input Single Output
SL	SideLink
SL-BCH	Sidelink Broadcast CHannel
SL-DCH	Sidelink Discovery CHannel
SL-SCH	Sidelink Shared Channel
SLA	Service Level Agreement
SLO	Service Level Objective
SLSS	Sidelink Synchronization Signals
SME	Station Management Entity
SM-NSSAI	Session Management - Network Slice Selection Assistance Information
SNMP	Simple Network Management Protocol
SNR	Signal to Noise Ratio
SNTP	Simple Network Time Protocol
SNV	Slice-specific Network View
SON	Self Organizing Network
SRB	Signalling Radio Bearer
SRS	Sounding Reference Signal
S-RSRP	Sidelink Reference Signal Received Power
SSM	Signal Strength Map
SSRS	Sidelink Sounding Reference Signal
SSS	Secondary Synchronization Signal
SSSS	Secondary Sidelink Synchronization Signal
ST	Selection Transmission
STA	Station

STCH	Sidelink Traffic Channel
SU	Secondary User
TAS	Transmit Antenna Selection
TBS	Transport Block Size
TDD	Time Division Duplexing
TFT	Traffic Flow Template
TLC	Transport Link Control
TM	Transparent Mode
TMD	Transparent Mode Data
TN	Transport Nodes
TRP	Time Resource Pattern
T-RPT	Time Resource Pattern of Transmission
TTI	Transmission Time Interval
TTT	Time to Trigger
TX	Transmitter
UE	User Equipment
UE-R	User Equipment Relay
UL	UpLink
UM	Unacknowledged Mode
UMD	Unacknowledged Mode Data
UMTS	Universal Mobile Telecommunications System
UP	User Plane
UPF	User Plane Function
USIM	Universal Subscriber Identity Module
V2V	Vehicle to Vehicle
V2X	Vehicle-to-X communication

VDSL	Very high rate Digital Subscriber Line
VOIP	Voice Over IP
vRP	virtual Radio Processing
WAS	Wireless Access Systems
WCDMA	Wideband Code Division Multiple Access
WiFi	Wireless Fidelity
WLAN	Wireless Local Area Network
WP	Work Package
WT	Wireless Termination
ZP	Zero Power

List of figures

List of tables

1 Introduction

Efficient management of wireless and mobile network resources is crucial for meeting the targeted performance requirements of 5G. Since one of the ambitions of 5G is to support very diverse services even with conflicting requirements, it is expected that next generation heterogeneous mobile networks (HMN) will be highly dynamic and even unstructured, at least for some applications. Such a big variety of services, coupled with a high heterogeneity of the physical infrastructure, calls for the development of management and control frameworks capable of acting on monitoring information and infrastructure variations (e.g. multi-tier with different cell sizes) in a scalable and efficient manner. At this aim, one concept intensively investigated in 5G is network programmability. In this framework, network function virtualization and software-defined networking are fundamental tools which can offer the desired flexibility of the system. They allow a simplified development and deployment of virtual network functions in opportunely identified nodes of the network. This approach can be adopted for traditional network control functionalities like load balancing, multi-connectivity, traffic-steering, and spectrum management.

Even if the study and proposal of control functions in this new framework is quite important, enabling network programmability in practice implies the existence of interfaces giving control functions access to underlying novel and legacy radio access technologies (RAT). A grand challenge is to provide, when possible and useful, a unified and RAT-agnostic exposure of networking conditions and control parameters via high-level, expressive abstractions across heterogeneous RATs. Dealing with complex control functions in heterogeneous networks today encompasses manipulation of infrastructure parameters based on physical (PHY) layer (e.g. channel, interference, power) and medium access layer (MAC) (e.g. data buffer, Layer 2 throughput) metrics and measurements, but these are highly specific to the RAT and vendor, requiring further aggregation processing towards RAT-agnostic representations.

This document contains the contribution of the COHERENT WP3 towards the objectives summarized in the big picture above. Contributions of the first area deal with the modeling and abstraction[1] of MAC and PHY layer status and behaviors with a particular focus on LTE-A, WiFi and in some cases the future 5G New Radio (NR), e.g. for the millimeter waves radio access. Ideally, it is possible to represent the status of a RAT with a data structure collecting all network elements, and all possible metrics between any pair of network elements. However, the COHERENT approach for representing a network status is function-driven: in the COHERENT project a network graph is an aggregation of the available information which is necessary to a specific control function. This document focuses then on metrics and network graphs for a subset of all possible control functions, namely load balancing, traffic steering, resource allocation problems in D2D communications including coverage extension, control of specific features of distributed and massive MIMO schemes, etc. Network graphs are tightly matched with the architecture assumptions described in COHERENT D2.2 [3] and with the control function using them. Nevertheless in the document we tried to highlight and stress in easily-identifiable subsections the findings on network graph topic.

Concerning modeling and abstraction of wireless networks, it is important to emphasize that there is (and there will probably always be) a need to build models for technology-specific techniques. In fact, very often, new techniques are proposed inside a given RAT, see for instance Device-to-Device (D2D) communications in LTE-Advanced Pro[2] (LTE-A Pro), or massive MIMO in the future 5G. As a first step, these techniques mandatorily need to be studied and modeled, before being able to integrate them in the wider picture of the model of a given RAT. In this deliverable

[1] An articulated definition of the term “abstraction” is provided in Sec 2.1 of COHERENT D3.1 [4].

[2] 3GPP uses the name LTE-Advanced Pro for referring to Releases 13 and 14.

the reader will find results of this kind, for instance the modeling of D2D communications in LTE-A Pro in Sec. 4.3.1, which, starting from MAC and PHY layer parameters of this technique, gives a model and a throughput metric in presence of collisions. Besides contributions on RAT-specific techniques, other results in this deliverable contribute to enlarge the view to technology-agnostic models of function-driven parameters. For instance, Sec. 4.4.1 presents attainable throughput as a RAT-agnostic network state abstraction over WiFi and LTE.

Finally, we wish to stress that the work toward providing a completely RAT-agnostic view of the wireless infrastructure is immense; the work done in COHERENT WP3 is a first step towards this ambitious goal. This is due to the fact that different RATs often have completely different approaches, e.g., in the management of the wireless media. Let us think for example of the scheduled access in LTE versus contention-based access in WiFi. Just to quote some other issues, we can mention the high heterogeneity of the parameters and parameter types among RATs and the identification of their "potential interest" (not all parameters bring gains in being abstracted to RAT-agnostic); the impact of vendor specific implementation of some parameters in terms of definitions and ranges; careful analysis of output or aggregated metrics to ensure that the observed network state or performance is comparable between different RATs (e.g. LTE and WiFi).

The second area, which is tightly coupled to the first one, deals with studies on examples of specific control functions. These studies have been performed in the framework of the COHERENT architecture which has been specified in D2.2 [3] and is recalled, with some extensions, in Sec. 2. As far as possible, constraints imposed by the COHERENT architecture have been taken into account in our investigations. For many contributions, we describe which part of the global control function should be implemented locally, in the so-called Real Time Controller (RTC), and which part could benefit from a logically-centralized implementation in the Central Controller and Coordinator (C3). Each control function is inserted in a monitoring cycle and a command cycle. The monitoring cycle is used to learn the environment, whose views are aggregated at different levels of the architecture in order to generate network graphs, as we have seen earlier in this section. The command cycle needs also interfaces for sending commands generated by the decisions taken by the intelligence of the control application. A wide set of applications have been evaluated either by inspection or, more often, by simulation, organized in four categories: load balancing, multi-connectivity and multicasting, relaying and D2D communications, which include coverage extension, and link performance, which includes results on MIMO schemes and inter-cell interference coordination. More details are provided in the next subsection of this introduction. Another important topic related to control functions are the interfaces, messages and procedures that can be used to convey information about control and coordination decisions. Practical examples and definitions of these messages and procedures are given in Subsections in 2.3, Section 5, and Annexes for systems and problems more closely related to standardization. Those sections show the strong commitment of the consortium towards standardization efforts, and they constitute part of WP3 output towards WP7.

1.1 WP3 Objectives

This section aims at summarizing the main objectives of WP3 and helping the reader to browse the document. Besides the traditional way to present the content of a technical report such as this deliverable, this section provides a so-to-speak "reading path" which relates the WP objectives with the corresponding deliverable sections. This may be particularly useful to readers interested only in specific topics.

The following list quickly summarizes the main WP3 objectives coming from the DoW:

[O1] Modelling and abstractions for inter-cell radio resource allocation, D2D communications, and mobility management.

[O2] Develop control applications for flexible inter-cell resource allocation and interference coordination.
[O3] Study and propose novel abstraction-based control methods for BS cooperation, e.g. for MIMO schemes.
[O4] Study D2D communications and new abstraction-based control methods Study user cooperation and propose new control methods for D2D communications.
[O5] UE and network mobility management via abstraction-based control.
[O6] Design a control protocol for LTE and 5G enhancements.

In the following table, the sections dealing with each objective are listed. Please note that certain sections may deal with multiple objectives.

Table 1-1: Mapping of main WP3 objective to sections of the document

Objective	Section
O1	All sections deal with this objective. For a summary of metrics see Section 3
O2	4.1.1: Load balancing in dynamic multitier heterogeneous networks 4.1.2: Load balancing using traffic steering in HetNet 4.1.3: Load balancing using probabilistic models 4.2.1: Multi connectivity in LTE 4.2.2: Multi-connectivity in mmWave networks and LTE 4.2.3: Multicasting in WiFi 4.4.2: Distributed antenna systems for throughput improvement 4.4.4: Inter-RAT spectrum resources sharing 4.4.5: Inter-cell interference coordination
O3	4.1.3: Load balancing using probabilistic models 4.2.2: Multi-connectivity in mmWave networks and LTE 4.4.3: Massive MIMO Exploitation to Reduce HARQ Delay in PHY/MAC
O4	4.3.1: Modelling of SideLink Communications in LTE-A 4.3.2: Coverage extension through D2D for LTE and evolutions 4.3.3: D2D cooperation and relaying 4.4.1: Building probabilistic network graphs with RAT-agnostic abstractions
O5	4.2.1: Multi-connectivity in LTE 4.2.3: Multicasting in WiFi 4.3.2: Coverage extension through D2D for LTE and evolutions 4.3.4: Mobility study for high speed platform 4.3.5: QoS Guaranteed Moving Relay in Self-Backhauled LTE 4.4.1: Building probabilistic network graphs with RAT-agnostic abstractions
O6	2.3.2: Heterogeneous IEEE 802.11 – 3GPP LAA network 2.3.3: Transport Link Control for entities in cellular deployments 2.3.4: Control framework for enforcing policy rules over CU-DU interface 2.3.5: User plane integration in centralized deployments 2.3.7: Control procedures for user plane operation 5: Control Interface APIs, Annexes

Table 1-2 presents the mapping of sections to their main feature, once again to help the reader browsing the document. Approximately one half of the contributions describe control methods which are applicable, at least qualitatively, in a multi-RAT scenario. WiFi and 5G systems are the RATs which are mostly coupled with LTE or LTE-A in case of multi-RAT control method. Moreover, most of the contributions describe control methods which involve the use of two levels of control, the first one is local and real-time (or time-constrained), whereas the second one is central and with relaxed time-constraints. The reader can refer to the contributions for details about the mapping of the control functionalities to the architecture entities. Notice that Sec. 4.3.1 which

is specifically devoted to modeling of D2D communications in LTE-A, contains no control application. Certain contributions were selected to provide an output to WP2 (as far as the Software Development Kit – SDK – is concerned) and WP6, i.e. demonstration.

Table 1-2: Mapping of subsections to main features of the contributions and to related WPs

	Description of the method		Control applicability			
Section	Single RAT	Multi-RAT	RTC (real-time/regional control)	C3 (logically-centralized control)	Feedback to SDK and demonstration	WP2 Use Cases
4.1.1	X		X	X		UC1.RS
4.1.2	X	qualitative	X	X	Yes	UC1.RS
4.1.3	X	qualitative	X	X		UC1.RS
4.2.1	X		X	X		UC3.FS, UC6.MR
4.2.2		X	X	X		UC1.SR
4.2.3	X		X		Yes	UC6.MR
4.3.1	X		NA	NA		UC3.CE
4.3.2	X		X (time-constrained)			UC3.CE
4.3.3	X		X	X		UC3.CE
4.3.4	X		X (time-constrained)			UC6.AG
4.3.5	X		X	X		UC3.MN, UC3.CE
4.4.1	X	X	X	X	Yes	UC1.SR Abstraction tool
4.4.2	X		X	X	Yes	UC2.MM
4.4.3	X	X	X	X		UC2.MM
4.4.4		X	X	X		UC2.FSA
4.4.5	X		X	X		UC2.MM

The last column of the previous table presents also a mapping of D3.2 contributions to main applicable Use Cases (UC) described in COHERENT D2.1 [2]. The key use cases covered are RAN sharing (UC1.RS), UC1.SR on supporting multiple RATs, UC3.CE on coverage extension and out-of-coverage communications enabled by D2D communications, UC6.MR on delivery of services in public or private transportation in urban areas and UC2.MM on Massive MIMO/distributed antenna system in dense small cell deployments. Note however that many studies are focused on control applications or tools for abstraction and modeling which can be applied in a scope larger than the one of the use cases of COHERENT D2.1 [2], quoted in the previous table.

1.2 Structure of the deliverable

After this short introduction, Sec. 2 provides a brief introduction to abstractions by network graphs which was presented in detail in D3.1 [4], then the section recalls the COHERENT architecture, which was defined in the scope of WP2. An extension of the architecture for D2D communications is also proposed, with some examples of applications of the monitoring and command/control cycles. Finally, in this section we report standardization-oriented investigations on the centralized control of different entities, in a cellular system or combined cellular and WiFi system, while applying COHERENT architecture. In particular a control framework in the 5 GHz band is proposed for coordinating an LTE system in unlicensed bands coexisting with a WiFi system.

Notice that the 5 GHz band is one of the few, if not the only, case in which LTE and WiFi are allowed to operate in exactly the same frequency band. This work has been presented at the ETSI BRAN and it is a standardization output of the project. Work related to contributions to 3GPP RAN3 on 5G new network architecture and interfaces is also reported. This is one output of WP3 toward WP7.

A summary of the metrics for network graphs is provided in Sec. 3. It highlights in particular the main RAT-agnostic abstractions that are used over the document. An effort of classification of such abstractions has been done and reported. Finally Sec. 3 collects also a list of novel expressions of high-level abstractions produced in COHERENT.

Sec. 4 collects the results either on modeling and abstractions of PHY/MAC layer parameters, or on the applications of abstractions to targeted control applications. Results are presented in groups of thematic areas rather than per WP task.

Contributions in Sec. 4.1 deal with load balancing from different perspectives, showing the benefit of having a centralized view of the network state. Different abstraction proposals are investigated, resource depletion probability, node capacity, overloads risk estimates. Load balancing can be jointly used with RAN sharing, where the load balancing is executed not only between physical base stations, but also between virtual base stations given the contractual agreement between virtual network operators that allows each virtual mobile operator to get access to the RAN of the other virtual network operator. In the absence of agreement between virtual mobile operators, the licensed operator can execute load balancing and distribute load in the network so that business agreements with virtual mobile operator will be not violated and the resources of the network will be used more efficiently. This will help the licensed operator to adopt more virtual mobile operators for RAN sharing or improve service for own customers.

Current and future heterogeneous networks pose challenges to mobility management, not only due to the ever-increasing requirements in terms of latency and throughput, but also due to the highly unequal size of the cells due to different technologies or simply to power transmission levels. Multi-connectivity is a tool for helping mobility management in such a context, and it has been investigated in Sec. 4.2. This challenge has been addressed by exploiting the COHERENT architecture, where the presence of a centralized coordinator enables mobility management different from the traditional handover, until the users are inside the region controlled by the coordinator. The investigated scenarios are focused on LTE and on 5G extensions of LTE with millimeter wave base stations. An improved way to control fundamental parameters of WiFi multicasting is also presented in this section.

Sec. 4.3 gathers the results on studies on D2D communications and relaying coming both from T3.2 which was focused on D2D communications under coverage, and T3.3 which studied out-of-coverage D2D communications, direct communication for coverage extension and the concept of moving cell. The first sub-section of this group deals with modeling of MAC behavior of D2D communications in autonomous mode (i.e. out-of-coverage) in the context of LTE-A Pro. Medium access in this mode of operation is based on contention and hence, behavior needs to the modeled and a weighted metric of throughput is presented. The following section presents how we can leverage on COHERENT architecture for implementing a coverage extension through D2D links and the concept of User Equipment Relay (UE-R). The third subsection considers in-coverage two-hop D2D relaying to enhance end-to-end throughput in multi-cell downlink networks, proposing an interference model that can be used by a central coordinator to determine some fundamental parameters of D2D communications and their activation. The fourth subsection presents the moving relaying concept in the context of air-to-ground communications. Here, the role of the controller is to find a reasonable configuration for SNR estimation under a significant Doppler shift experienced by the receivers. Finally, a fresh look to the moving cell concept is provided by the last contribution of this section. Here, existing LTE features are combined together in an

innovative way to create a wireless mesh. The concept of the COHERENT controller helps a lot in tuning the resource allocations between infrastructure links and access links.

Sec. 4.4 contains results about techniques dealing with PHY layers technologies like distributed antenna systems (DAS) or massive MIMO. The first subsection deals with load balancing via traffic steering in dynamic multi-tier heterogeneous networks: it is shown how it is possible to connect several parameters into single probabilistic abstraction in order to model the stochastic behavior of the resource consumption of base stations of different tiers in the cellular network. Sec. 4.4.2 describes a DAS system model and derives the probability of coverage metric for analyzing different DAS scenarios i.e., varied number of RRHs in a cell and targeted SINR at different UE positions. The third subsection shows how the abstraction and network graph can be used for new technologies like massive MIMO within COHERENT architecture and for heterogeneous mobile networks, for lowering delays in case of time-constrained applications. The fourth contribution proposes an inter-RAT scheduling algorithm to share limited time-frequency resources among users utilizing various RATs, namely 2G and 4G in this case. The proposal can be of high importance in dense heterogeneous network utilizing limited frequency resources. Finally, this subsection ends with an evaluation of an Inter-Cell Interference Coordination (ICIC) algorithm which takes advantage of a centralized network view and of the COHERENT control architecture.

Sect 5 deals with the control interface. It presents the procedures used for the API used to define the messages between the C3 and the relevant network entities. The general approach for message definition follows the 3GPP approach as used in X2 interface defined in LTE standard and all the other interfaces defined in 3GPP RAN3, including interfaces that will be used for 5G New Radio. This is another output of WP3 toward WP7.

The main conclusions of the report are collected in Sec. 6.

Annexes contain material pertaining to Sec. 2 and Sec. 5 like the detailed definition of messages, information elements and procedures, which were not judged necessary for the understanding of the investigation goals reported in the main text of the document.

Finally we recall that, in order to manage the length of the main text of this deliverable, detailed development and results of certain studies are reported in the following reports which are available to the reviewers: COHERENT M3.2 [5] and technical report “Modelling of D2D communications in LTE-Advanced Pro” [9]. Moreover, in order to avoid duplication, some contributions in D3.1 [4] are not included here. Thus, this document is not stand-alone, but is intended to be attached to D3.1, as they present together the technical output of this work package.

2 COHERENT Framework

This section is devoted to work on the control architecture proposed by the consortium. The first subsection reviews basic facts about network graphs, i.e. network views presented to the control applications running at different levels of the control plane. The concept of offering a compact representation of the network state to a specific function is, of course, not new. Here, however, we propose to explicitly integrate it as an element of the global architecture. COHERENT network architecture, presented in COHERENT D2.2 [3] is recalled in Sec. 0 Sec. 2.2 presents an extension of the same architecture for D2D communications, and also gives high-level descriptive examples of the monitoring and command cycles in that case. Section 2.3 is an output of WP3 towards WP7: it deals with adaptations of the COHERENT architecture to standardization efforts in different scenarios. We report the main points of the output produced for the ETSI BRAN about a central coordinator and controller for efficient management of LTE-LAA (Licensed Assisted Access), WiFi and 5G (New Radio) coexistence in the 5 GHz band. Then, part of the output for 3GPP standardization work is presented, on different topics related to with the new architecture including Central Units (CU) and Distributed Units (DU) standardized by 3GPP. Shortly, the functionalities and protocol stacks of an eNodeB are split between one central entity (CU) and one or more remote entities (DUs). We note that the CU/DU split under definition by 3GPP is a first step in the direction of the COHERENT architecture.

2.1 Background: Abstraction by network graphs

Relevant background and challenges, which affect the work in WP3, are described with extensive literature review and reference list in D3.1 [4]. In this section, we only briefly summarize some main points presented in section 2 of D3.1 with some updates induced during the conducted work.

Network graphs (NGs) play a key role in modelling and abstracting the properties of wireless networks. A classic example of the use of NGs is the frequency assignment operation modelled as a graph colouring problem. Furthermore, NGs can be used to model connectivity, channel quality, and interference when multiple transmitters employ variable channel widths.

Control applications largely define the requirements for abstraction of control information. The current LTE-A cellular system must perform a number of radio resource management (RRM) tasks, such as user association between tiers and radio access technologies, resource allocation, mobility management, routing, and link adaptation. The objective of RRM techniques is to ensure sufficient quality of service (QoS) for individual users while maximizing the resource utilization of the network without causing significant cross- or co-tier interference to other collocated users. Moreover, in modern multiservice networks with heterogeneous QoS user demands require some fairness and prioritization capabilities from the RRM. Some RRM functions are executed during the setup of new connections while some functions are conducted for every packet arrival.

Abstractions of various phenomena have been extensively used for different purposes in the design of wireless networks. Regarding the above-mentioned control applications, the abstractions can be used to simplify the controlling approach with the aim to reduce the control overhead and complexity of required control methods for RRM. Simplification is emphasized in systems using centralized control. An outline of abstraction concepts in wireless networks is provided in Figure 2-1.

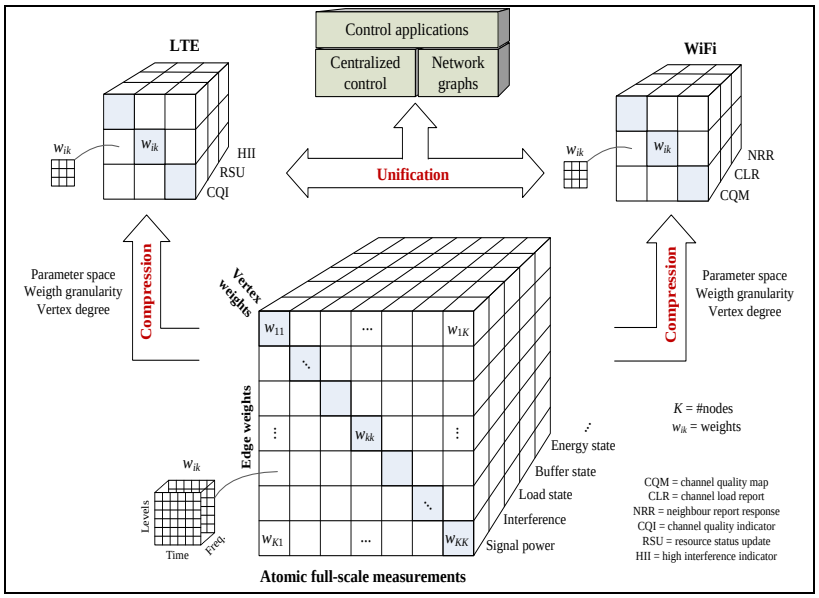


Figure 2-1 Illustration of abstraction concepts in wireless networks using centralized control

The main abstraction tools toward simplified control approaches are i) compression and ii) unification which take atomic full-scale measurements of the network state as inputs for further processing. Both concepts are important in the overall abstraction framework but with different emphasis. Compression means that unnecessary information is removed or hidden possibly behind another parameter that can be more efficiently conveyed in the control channel. Examples of compression factors, which can be affected, are network parameter space, NG weight space-time-frequency granularity, and NG vertex degree defining the interaction between network nodes. The compressed control signals need to be unified, so that a centralized common control unit can interpret control information correctly for all involved radio access technologies, such as LTE and WiFi.

COHERENT network architecture

Figure 2-2 shows the functional blocks for implementing the COHERENT architecture. Two control/coordination mechanisms namely C3 and RTC provide scalability and time-wise adequacy for control and coordination.

By receiving, through the southbound interface (SBi), status reports from lower layer entities, C3 maintains a centralized network view of the governed entities, e.g., transport nodes (TNs) and Radio Transceivers (RTs) in RAN. The acronym RT is referred to as a generic element in the Wireless Access Systems/ Radio Local Area Network (WAS/RLAN). For example, an RT could be a legacy LTE eNB or a legacy WiFi AP or a New Radio Base Station (NRBS). The network entities (TNs and RTs) connect to C3 and RTCs through the southbound interface.

A network slice in the service plane is defined as a collection of specific network applications and RAT configurations. Different network slices contain different network applications and configuration settings. Through the northbound interface (NBi), C3 and/or RTCs provide the required network view, namely slice-specific network view (SNV) for the network service slices so that network service slices could express the desired network behaviours (by programming) without being responsible themselves for implementing that behavior (with hardware).

Some application modules in network slices may be latency-sensitive. For such slices, these modules are located in the RTC. Examples of latency-sensitive functions are RAN function splitting in 3GPP New Radio, MAC scheduling, HO decision, MAC/PHY reconfiguration.

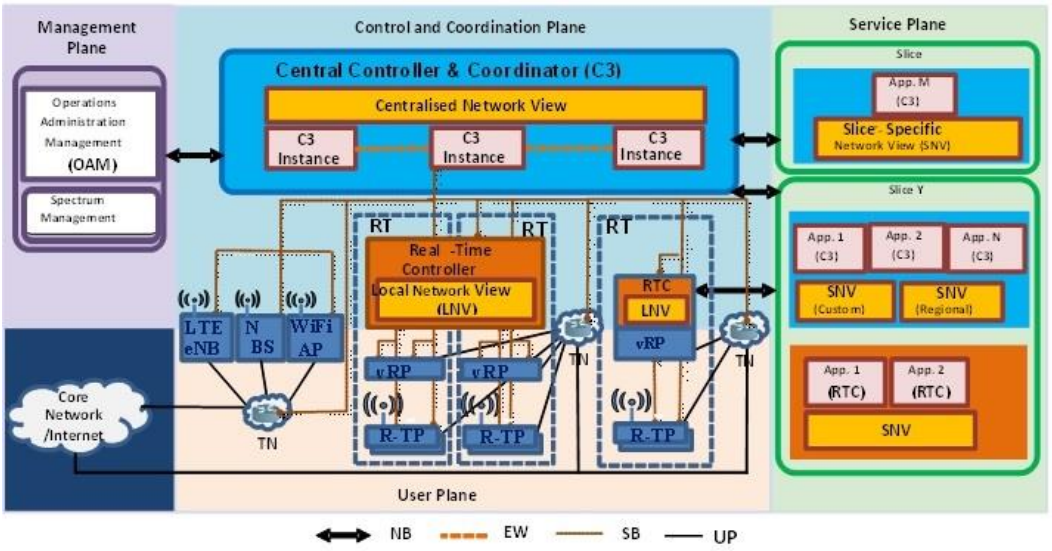


Figure 2-2: Functional blocks for implementing the COHERENT architecture (From D2.2)

In general, different NVs could be inside the same slice, according to what the application wants to do. While the control and coordination entities on the control and coordination plane make control decisions for RAN functions and send the decisions to the network entities for executing the decisions, the management plane usually focuses on monitoring, configuring and maintaining the long-term decisions for network entities in the infrastructure. The entities in management plane are connected to C3 through NBi. Further details about COHERENT architecture can be found in D2.2 [3].

The network representation presented in the previous subsection is in the form of a graph where vertices represent for example eNBs and edges represent the link's delay, utilization, energy consumption, etc. The collection of data is done by Network Information Functions (NIFs) inside the TNs, which are located in the control plane, and access and collect data from network elements and may process them to obtain a more convenient representation of the network status. The control plane, Central Controller and Coordinator (C3), is the network's brain which is responsible for monitoring the network, making decisions on control functions running on it, delegating control functionalities with stricter time-constraints to local control entities named Real-Time Controllers (RTC) and programming the behavior of the virtual and physical network. The NIF manager is the element that provides information about the NIFs. The abstraction layer consists of network representations in the form of network. The service plane takes the necessary network graphs as input, runs algorithms in order to minimize a set of criteria linked to the services, like data delay, link over-utilization, load balancing criteria in different network legs and returns a set of rules to the control plane.

2.2 Control framework for D2D communications

This section is an output concerning the system architecture of the COHERENT control plane studied in WP2, and can provide useful ideas of spectrum management related to D2D.

As recalled in Sec. 0, the COHERENT project proposes a flexible control plane for next generation wireless systems, including different Radio Access Technologies (RATs) and physical transceivers (Access Points, Base Stations, User Equipment, Relays). It leverages on the concepts of virtualization and Software Defined Networking (SDN) in order to improve the flexibility in the placement of the control functions. It uses abstracted network views in order to simplify its coordination functions over multiple RATs and, at the same time, to provide a simplified view of

the network state to a given control application by hiding a subset of the lower layers details which are not useful for that application.

COHERENT D2.2 [3] presented the general concept of the COHERENT architecture as well as a first version of it to Device-to-Device (D2D) communications[3] and relaying. Here we provide more details about the architecture in the D2D case and in particular on relaying based on D2D communications.

Every control function in a control plane works according to a cognitive cycle in which, very schematically:

1. information is retrieved from the system (reporting),
2. decision is taken by the control function or application or plan (decision step),
3. Enforcement of the decision on the system is done (command) in order to modify its behaviour.

The decision step mainly depends on the available information and the specific algorithm (see section 4.3 for examples of control applications). The architecture has an impact also on the reporting and command steps, as well on the placement of the decision function. The control plane configures the reporting procedures so that, for instance, more or less information become available trading it off typically with a reduction of the resources available for data transmission.

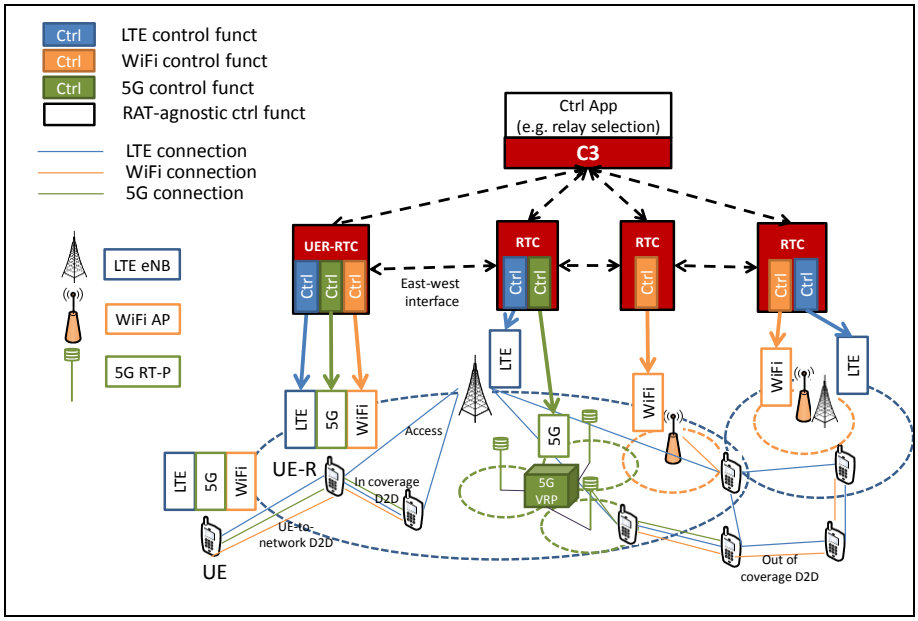


Figure 2-3: Example the extended COHERENT architecture with D2D communications

A new UE-Relay Real Time Controller (UER-RTC) is added in the control plane as a new optional entity. This UER-RTC is a Real Time Controller (RTC) instance located in the UE-Relay (UE-R) which is typically in coverage, e.g. like in Figure 2-3. The functionalities of the UER-RTC are not

[3] D2D communications in the context of the evolution of LTE inside 3GPP standard take different names like LTE-Direct, Proximity Services (ProSe), Sidelink communications, depending on the type of the topic treated (respectively marketing/commercial use, services, low layer technical specifications). D2D communications in the context of IEEE 802.11 can be implemented through WiFi-Direct or extension 802.11s.

different from the ones of an RTC, and it is in charge of real-time control tasks (like resource allocation, interference coordination, relay selection, etc.) only on a local and limited set of links, namely:

- In-coverage D2D relay links of which it is the source or destination. Two cases are possible, in-coverage in-cell or inter-cell.
- Out-of-coverage D2D links, even in the multi-hop case, attached to the network through a UE-Relay. A possible placement of the UER-RTC in this case could be in the UE-R being the gateway between out of coverage users and the network.

In 3GPP LTE-Rel 13, as an example, the eNodeB may control at least in part in-coverage D2D links. An important design constraint is that the control over a link should be uniquely defined. We suppose that C3 delegates control functions to RTC and to UER-RTC. It will decide for instance, which links will be controlled by the eNodeB or the RTC controlling the eNodeB and which links will be controlled by the UER-RTC located at the UE-R. This delegation is specific to a given control function. For instance, for resource allocation, three possibilities are available for in-coverage D2D links:

- Fast resource allocation is controlled by the RTC controlling the eNodeB. In 3GPP it corresponds to network-controlled allocation.
- Fast resource allocation is controlled by the UEs. In 3GPP, it corresponds to autonomous resource allocation, which works both in- and out-of-coverage.
- Fast resource allocation is controlled by the UER-RTC. In this way it is possible to implement a hybrid network-controlled resource allocation (not yet existent in the standard), because it is implemented by the UER-RTC located at the UE-R and hence messaging information flow may change.

Concerning out-of-coverage D2D communication limited control is proposed in LTE Rel 13. With the current proposal there could be a control over those links through the UER-RTC or the RTC associated to an eNodeB. In general, in the COHERENT framework we can even imagine that UER-RTC runs on UEs which are out-of-coverage, thus implementing a form of control in multi-hop scenarios. These UER-RTCs running on out-of-coverage UEs shall be able to keep a consistent network vision with the C3, which could be challenging in terms of signalling burden in a multi-hop context. For this reason, we expect that the concept of control delegation of the C3 to the UER-RTC could be even more heavily used for out-of-coverage UEs. Finally, if the connection between part of the out-of-coverage UEs and the network is lost, due to strong mobility or link quality degradation, then the UER-RTCs of the disconnected set of users may take up certain control tasks of the C3 in order to reconfigure and run the disconnected set. For this concept, please see also the work done in COHERENT WP5 D5.1 [6], and D5.2 [7].

Even if the UER-RTC is an instance of RTC and has then the same logical functions with a more limited scope, we assume that it has specific characteristics:

- It is physically instantiated on a UE-Relay (UE-R). As a consequence, the quantity of metrics and signalling, especially for out-of-coverage UEs, can be controlled and it has to be since the east-west interface with other RTCs always goes through a wireless link. Moreover, control functions should not be greedy in terms of computation and memory requirements.
- RTC may support multiple RATs, since usually current UEs are multi-RAT. Whether one RTC should implement multiple-RATs or there should be one RTC per RAT, can be seen as an implementation detail, since the instances are collocated in the UE-R. In the next figures we have put one RTC over multiple RATs just for illustration.

- UER-RTC typically contains a subset or a light version of control functions running on an RTC instance associated to a VRP (or an eNodeB, WiFi AP), due to performance constraints of the platform.

In general, concerning a multi-RAT, RTC instance (not necessarily a UER-RTC), we assume a controller-agent software architecture.

The controller is the intelligence taking decisions in real time inside the RTC, and can delegate some decision and control functions to the agent. It makes the interfaces with applications (if they require directly real-time features), with the C3 instance, or with other RTC instances. It can be physically collocated or not with the device.

The agent is always physically collocated on the devices (eNodeB, WiFi AP, new 5G base stations, UEs, etc.) The agent is RAT-specific and can run some control functions as the delegation of the controller as per policy specified by the controller to make time-critical decisions.

A schematic view of the controller/agent architecture is shown in Figure 2-4. It is software architecture, rather than a functional architecture. As a matter of fact, the agent is a piece of code which can host part of the decision algorithms on behalf of the controller. The agent is collocated physically on the device and its device/agent Application Programming Interface (API), which is the southbound interface of the COHERENT control plane, interacts with both the control and user (data) plane of the standard implemented in the device. Hence, the RTC can directly and dynamically configure the user plane or choose to configure it through the control plane of the implemented standard. Moreover, for instance in 3GPP protocol suite, even if signalling and control messages are defined, certain control functions are not defined in the scope of the standard, for instance the Medium Access Control (MAC) scheduler in real time. The COHERENT control plane provides also a framework to deploy those functions in a flexible way. Notice that in the figure the box with the standardized protocol stack does not represent the physical device, but only the software implementing a standard protocol stack (WiFi, LTE, 5G, etc.). Moreover, especially for already existing standards, part of the Control Plane (CP) and User Plane (UP) may be mixed in particular at low layers of the protocol stack. For instance, in LTE in the PHY layer, control messages can use the UP data flow for being transmitted.

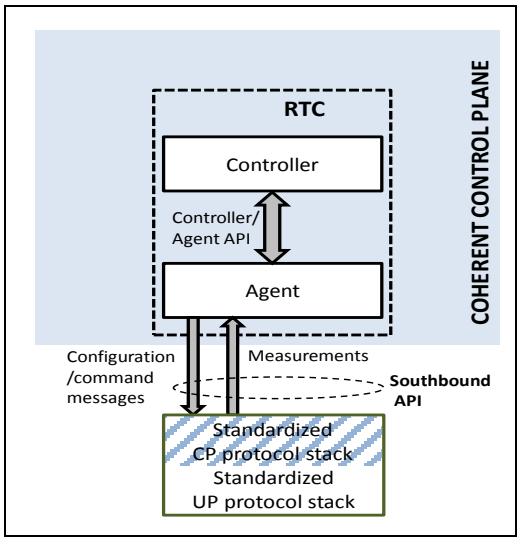


Figure 2-4: General controller/agent architecture for Real Time Controller (RTC)

Coming back to Figure 2-3, it shows an example that includes D2D communications in COHERENT architecture. Vertical boxes over the physical devices (UE, eNodeB, AP, 5G-VRP, etc.) show the protocol stacks of existing or new (5G) standards which are instantiated in hardware (those boxes are not shown for all physical devices for readability's sake). The links provided by the different RATs are also shown by lines in between the devices (illustrative example). A control application (or function) can run in the C3, and in the figure this control application is RAT agnostic, as an example.

The C3 control application may imply that different RAT-specific control functions are deployed and run on all or part of the RTCs connected to the C3. An example of this configuration will be given, for instance in section 4.2.2 for the multi-connectivity case, but without D2D communications.

Control functions can be deployed also on the UER-RTC, at the very edge of the network. This can help real time tasks, for D2D links which are out-of-coverage (but connected to the network), because it reduces the number of hops and the signalling burden over the RAN (no wired connection between the UER-RTC and the other instances of COHERENT control entities). An example of this situation is given for instance in section 4.3.2, where coverage extension through D2D is studied.

There may be no need to support multiple RATs. For instance in section 4.3.3, C3 controls multiple Radio Transceivers (RTs), which in this case are LTE eNodeBs. The RTC of the eNodeBs are in charge of fast resource allocation for standard and D2D links, while UEs takes care of direct communication and of gathering measurements. This example is a case of the more general topic of interference management for RT-controlled D2D communications. Another class of applications which can take advantage of a centralized coordination is the coordinated scheduling of links which are not in the same cells. Section 4.3.5 shows an example of application over a wireless mesh. In this case, direct links are not implemented through D2D but thanks to an original relaying scheme.

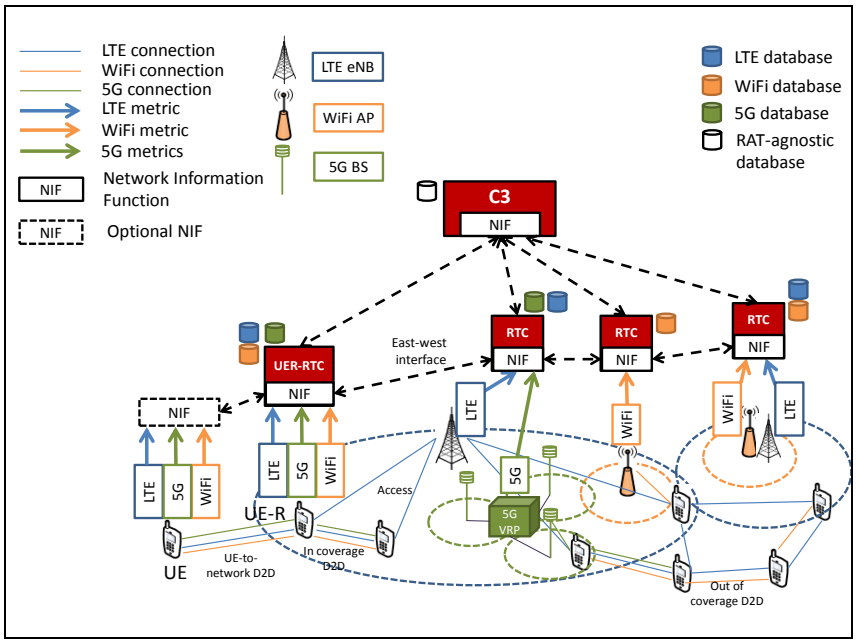


Figure 2-5: Example of reporting flows in the COHERENT architecture with D2D

Figure 2-5 gives a schematic representation of the COHERENT architecture for the Radio Access Network (RAN) part and it stresses in particular monitoring flows.

The COHERENT architecture uses Network Information Functions (NIFs) as defined in D2.2 [3], which are used for monitoring and information sharing. Examples of NIF functionalities are:

- Collecting the measurements (per RAT).
- Processing the measurements (e.g. probabilistic models for metrics).
- Manage measurements or metric transfer (for instance in case one RAT is failing).

NIFs can be seen as structural functions for implementing reporting/messaging internal to the COHERENT control plane and they can work at a different level of abstraction. Notice that, in the controller-agent architecture, NIFs may also be instantiated on the agent side on the physical device. In that case, local NIFs work on the measurements of a given RAT and can be used to process measurements (like filtering the number of measurements passed to the RTC, or erasing the outdated measurements) for reducing the signalling burden between the agent and the controller.

In the figure optional NIFs are also present on out-of-coverage UEs. These local NIFs could be useful for reducing the signalling burden in case of multi-hop networks, if some reporting is required by the UER-RTCs or RTCs.

Concerning network graphs, Figure 2-5 gives an example of technology-agnostic network graph in C3, while RTCs (including UER-RTC) have technology-dependent databases, meaning that the NIF in C3 is doing the translation to achieve the required level of abstraction. However, it is not the only possibility. Also RTCs could have RAT-agnostic network graph if necessary for the control function running on them.

2.3 Control framework for heterogeneous networks

2.3.1 Introduction

This section is focused on the centralized control of different entities, in a cellular system or combined cellular and WiFi system, while applying COHERENT architecture, addressing the following topics:

- Networks including entities (AP, eNB) based on IEEE 802.11 and 3GPP LAA;
- Backhaul integration within cellular entities, based on Transport Link Control function;
- Integration of system QoS.

The material of this section has been used in contributions to ETSI BRAN on Central Control of WiFi and LAA networks (draft TR 103 494) and to 3GPP RAN3 on 5G new network architecture and interfaces.

2.3.2 Heterogeneous IEEE 802.11 – 3GPP LAA network

2.3.2.1 Motivation and goals

Starting with 5 GHz and possible other frequencies like 3.5 GHz in US, the IEEE 802.11-based and 3GPP LTE-based systems can operate in the same band or even in the same channel. However, there is *n*o common protocol for management and distributed or centralized control of the nodes belonging to both technologies for achieving high spectral efficiency and user experience.

The coexistence of the different systems in these bands is based on a mechanism named Carrier Sense Multiple Access with Collision Avoidance (CSMA-CA). In 802.11, there is a PHY-independent Clear Channel Assessment (CCA) sense named CCA-ED (Energy Detection), based on a predefined threshold, and an 802.11-specific CCA, set at the sensitivity level. In 5 GHz there is a significant gap, i.e. 20dB, between these thresholds. The Virtual Carrier Sense mechanism allows sending a message for medium reservation to the neighbour nodes.

While 802.11-based system will detect another 802.11-based system in a 20MHz channel at the sensitivity level of -82 dBm, an LTE system will be detected at the much higher CCA-ED level (-62dBm for a 20MHz channel). This fact will strongly limit the coverage and throughput of an LTE-based system, because when the channel will be assessed as free, a new STA transmission will take place, creating thus interference to the existing LTE transmission.

On the other side, LTE (actually the Release 13 version named LAA – Licensed Assisted Access) operates with energy detection levels set at -62dBm/20 MHz for a transmitted power of 13dBm (similar to WiFi terminals) and at -72dBm/20MHz for a transmitted power of 23dBm (similar to WiFi AP), so an LAA AP will cause a lower negative impact on the WiFi systems on the same channel.

The goal of this research is to define mechanisms and messages allowing LTE and WiFi operation with higher cell size and throughput, while using the C3 services. For example, these messages could allow dynamic threshold adaptation, thus mitigating the previous issue

2.3.2.2 Architecture for heterogeneous WiFi-LAA operation

The considered system is illustrated in Figure 2-6. It includes the following basic elements:

- A Central Controller and Coordinator - communicating with a 3GPP base station using a X2/Xn-like protocol and with an AP using SNMP or a X2/Xn-like protocol, the Central Coordinator has the role of coordinating the operation of the system to improve one or more performance indicators;
- An LTE–based base station using at least one frequency band in a shared spectrum, for example 5 GHz;
- A New Radio – based base station using at least one frequency band in a shared spectrum, for example 5 GHz;
- An 802.11 – based Access Point (AP) using at least one frequency band shared with the LTE or New Radio base station;

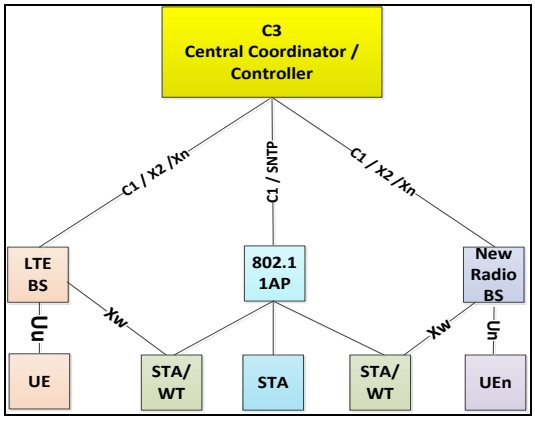


Figure 2-6 System description

- A UE (User Equipment) communicating with the base station using an air protocol developed by 3GPP;
- A station (STA) communicating with the AP using the 802.11 protocol.

A STA which is also a WT (wireless termination) and delivers data to a cellular subscriber; the WT is connected to the cellular base station through a protocol Xw developed in 3GPP.

A base station using cellular technology or Access Point using IEEE 802.11 technology may be implemented with the higher protocol stacks on a computing platform and with the lower protocol stacks on physical units including radios and antennas.

2.3.2.3 Monitoring of WiFi and LAA combined deployments

Existing messages in IEEE 802.11 standard

The IEEE 802.11-2012 standard defines several radio measurements that are reported by a station (STA). A Station Management Entity (SME) may receive the results of the radio measurements and set parameters of the STA. The SME, using an SNMP protocol, can further communicate with a compliant management entity.

The IEEE 802.11-2012 standard defines a high range of measurement reports, including PHY measurements such as:

- Frame Report, providing the measurement of average Received Channel Power Indicator (RCPI) which is an improved measurement as compared with RSSI;
- Channel Load Report, providing an indication of channel busy time;
- Noise Histogram Report, providing the Idle Power Indication as a function of time in idle intervals;
- Transmit Stream/Category Report, providing QoS related measurements, such as per traffic categories, where a traffic category is associated with a priority indicator.

Annex B provides a detailed selection of the IEEE 802.11 standard messages, which can also be extended and used by COHERENT C3.

Existing messages in 3GPP LTE standards

The 3GPP - LTE standard TS 36.213 v14.0.0 defines CQI measurements and LAA behavior, TS 36.214 v14.0.0 [13] defines the RSSI measurements to be reported by an UE, while TS36.423 v14.0.0 defines protocols for distributed coordination between base stations. All of them are designed for licensed bands but this does not preclude the usage of some of them in the shared spectrum as well.

2.3.2.4 Reports and Information Elements

Annex A captures the reports and information elements relevant to this section.

2.3.3 Transport Link Control for entities in cellular deployments

2.3.3.1 Motivation and goals

As part of the 3GPP 5G study [15], a base station architecture in which the base station includes a Central Unit (CU) and Distributed Units (DU) was proposed. Within the general COHERENT architecture CU is in fact the combination of C3 and vRP and DU is in fact the R-TP. As the CU-DU split is one of the 3GPP topics closest to COHERENT vision, it has been considered for studies here.The traditional protocol layers are split between CU and DU.

One of the drawbacks of this solution is that packets might be lost on the CU-DU interface without the possibility to identify the packets lost over the transport network. The goal of this section is

hence to introduce Transport Link Control functionality, with related messages, which is able to detect and correct the errors of the Transport Network (TN) between the CU and DUs. TLC impacts on QoS enforcement, slicing and mobility among different DUs. An evaluation of latency reduction is also provided.

2.3.3.2 Architecture for backhaul integration in centralized deployments

The system deployment is depicted in Figure 2-7. The BS is split between a CU and multiple DUs. The CU contains the UP part of the BS functions, for example the big group of RRC, PDCP, while the DU contains the rest of functions, for example RLC, MAC and PHY.

C3 may be a stand-alone unit on the control plane or may be part of the CU. C3 includes the RRM (Radio Resource Management) control functions (see TS 36.300 [12]).

The BS is connected to a Serving Gateway (SG) (in LTE) or to the UPF (User plane function) on the User Plane.

The BS may be also connected to a Mobility Control Entity (MME in LTE) or a Core Access and Mobility Management Function (AMF) in 5G on the control plane; this connection is shown in Figure 2-7 as going over the Transport Networks TN.

A Transport Link Control (TLC) function is implemented on each traffic direction. In the existing implementations this function is eventually implemented within the transport network, but in this document this function providing error detection and correction for the transport network is implemented within the cellular system.

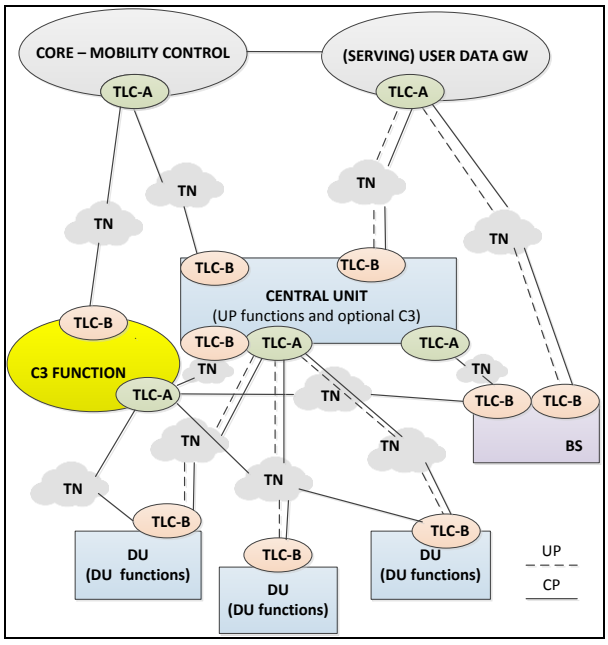


Figure 2-7 Architecture for backhaul integration in centralized deployments

TLC entities are hosted by the cellular entities sourcing the traffic; if the traffic is in downlink direction, the TLC at the cellular entity sourcing the traffic is shown in Figure 2-7 as TLC-A, encompassing all the entities of traffic sourcing TLC, while the pair TLC entity including all the TLC entities receiving the traffic is noted TLC-B.

Figure 2-7 shows two types of base stations, which can use the LTE or 5G technology: a BS which is split between a CU and several DUs and a BS which is not split. TLC operation is similar to RLC operation described in LTE standards.

Figure 2-7 illustrates two types of C3: one which is included in a BS, like C3 in the CU and one which is external. In a given deployment, only one variant should be used.

We consider a transport network between two cellular entities as being "protected" if there is at least one instance of a TLC entity implemented in a first cellular entity assigning a sequence number (Unacknowledged/Acknowledged Modes, UM/AM) and an instance of a pair TLC entity implemented in the second cellular entity.

One or more links of the TN can use protected transport, while in other case can be used the un-protected transport (Transparent Mode or nothing).

In the system deployment for uplink links the transmit and receive sides of the transport network links are reversed relatively to Figure 2-7.

2.3.3.3 RLC and PDCP functions description

The RLC and PDCP functional description is included in Annex A.

2.3.3.4 QoS enforcement

The 3GPP system enforces QoS by defining "bearers" in LTE or "5G QoS flows" in 5G. For each bearer in LTE a QCI identifier is defined which is a scalar identifying a QoS class; the QCI is mapped to a DSCP (Differentiated Services Code Point) value in IP header.

In 5G, as indicated in 3GPP TS 23.501 [10], all traffic mapped to the same 5G QoS Flow receive the same forwarding treatment (e.g. scheduling policy, queue management policy, rate shaping policy, RLC configuration, etc.). Providing different QoS forwarding treatment requires separate 5G QoS Flow.

In a similar mode, 5G QoS Indicator (5QI) is a scalar that is used as a reference to a specific QoS forwarding behavior (e.g. packet loss rate, packet delay budget) to be provided to a 5G QoS Flow.

2.3.3.5 Service Slice

The new concept of Network Slice enables the allocation of resources per service slice. For example, in TLC AM or RLC AM in this document, slices related to IoT service may require a higher reliability as compared with other services.

As mentioned in TS23.501 v0.2.1 (2017-01), "Network Slice Selection Policy (NSSP): This policy is used by the UE to associate UE applications with Session Management- Network Slice Selection Assistance Information (SM-NSSAIs) and to determine when a new PDU session should be requested with a new SM-NSSAI."

2.3.3.6 TLC support for QoS and Network slicing

In our system, an instance of the transmit side of the TLC entity shall be defined only for one combination of QoS and Slice identity, i.e. for each identified QoS category by QCI or by a QoS flow identifier and each identified Network Slice (NS) there shall be a separate instance TLS function.

In another approach, an instance of RLC or TLC entity can be defined per user application, i.e. per PDU session.An instance of TLC entity can be defined per network slice or per operator.

In case that the TLC is collocated with the source of the traffic, is possible to map the user or control traffic belonging to a specific QoS category and NS (if used) to a specific queue.

In case that the TLC are not collocated with the source of the traffic, the user data packet shall carry the information of the type of the traffic as applicable (user traffic, control traffic), control channel identifier if appropriate, specific QoS category and the identity of the Network Slice or of the PDU session. This can be done by inserting at a well-known position in the packet of an Information Element indicating as appropriate the QoS category or a NS identity or a PDU session identity or a Control Channel ID.

Each instance of TLC, especially if not co-located with the source of traffic, shall have its own identifier which will enable the differentiated configuration of each TLC instance.

A TLC instance or a function able to create different queues must be able to identify the appropriate user or control traffic based on appropriate filters (deep packet inspection) which can find specific values for QoS category or a NS identity or a PDU session identity or a Control Channel ID and direct the selected traffic to the appropriate queues or RLC/RTC instances.

For example, if a PDU session is identified by the scalar M and the QoS flow is identified by the scalar P, only those packets matching these values will be processed by the TLC instance.

2.3.3.7 Configuration of the TLC instance

The configuration of the TLC instance which is not collocated with the configuring entity will be done by signaling messages belonging to the Control Plane and generated by the RRC entity or C3 indicating the TLC instance identifier and configuring. The details are provided in Annex A.

2.3.3.8 Monitoring of backhaul operation

For allowing a control and/or coordination function or entity to make decisions about the TLC operational rules, each TLC instance shall provide reports to C3.

Such reports can include, for a defined interval:

- traffic amount
- number of lost packets
- packet or bit error rate
- average delay
- maximum delay for error correction through retransmissions using the Automatic Repeat reQuest (ARQ) mechanism
- number of PDU sessions having 1% or more lost packets
- Percentage of lost packets for a specific slice.

The reports can be provided per link, i.e. for all instances of TLC belonging to a specific link. For example, CU communication with DU takes place over a downlink link and over an uplink link. For each link and over a defined time interval it can be provided, preferably by the TLC receiving side, statistics for each UM and AM including the traffic amount, the number of lost packets or the packet error rate or bit error rate or average delay or maximum delay for error correction through retransmissions using the ARQ mechanism.

The above reports can be used by C3 in handover or load balancing for base station or DU selection, as the overall link behavior is influenced by both the air interface and the transmission network.

Control messages

To each TLC entity operating in one of the TM, UM, AM it can be assigned a specific TLC instance, by sending a configuration message or by a control message.

A TLC instance can be controlled for changing its mode of operation, by configuring one TLC entity defined for an operation mode to include the specific TLC instance.

A TLC instance can be instructed by a message to add more possibilities to the traffic filters, for example more QoS flows or more PDU sessions or to modify the existing packet filters by deleting some of the filter parameters, i.e. by removing from the filtering operation some QoS flows or PDU sessions.

2.3.3.9 Downlink UE mobility between DUs connected to the same CU

When the handover (HO) decision is taken by C3, it is needed to forward the PDCP packets to the destination DU instead of forwarding them to the source DU as before the HO.The anchor point when the delay of the transport network is small is the PDCP layer, which will stop sending the packets to the source DU and will route them to the destination DU.

It is possible to take advantage of the architectures in Figure 2-7, as described below. In all the options described below, the RLC is informed by the C3 about the HO decision for an UE or a PDU session of the UE or a QoS flow of the UE and about the destination DU IP address or identifier.

2.3.3.10 Congestion detection

TLC transmitter is placed within the CU or DU cellular entity and the TLC receiver is placed respectively in DU and CU entities.

In AM the transmitter receives by ACK/NACK the feedback on the success of transmissions and can provide, for a defined time interval, the number of lost packets or their percentage.

The time synchronization between the CU and DU is possible through different IP protocols or by satellite systems such as GPS. If a time stamp is added to the TLC PDUs, the receiver can determine the delay for each packet and create for a time interval a statistical representation (average, median, minimum, maximum, percentage of packets with the delay lower than a value) for the recorded delays. This representation can be done per QoS flow, per Slice or per UE PDU session. Based on the delay statistics and on the lost PDU statistics transmitted through signaling messages over the DU-CU interface to a C3, C3 can determine the congestion situation for a specific CU-DU link and decide to use another DU for part or all the UE-generated traffic. The delay statistics and the lost PDU statistics can be also transmitted through signaling messages within the CU.

For implementing the avoidance of congested traffic network C3 shall send a signaling message to the UE and the PDCP layer in the CU for executing the HO of the UE or of a part of its traffic such to avoid using the congested CU-DU links.

2.3.3.11 TLC impact on latency reduction

The delay in the high layer split (for example split Option 3) and in TLC approach can be deducted based on

Figure 2-8.

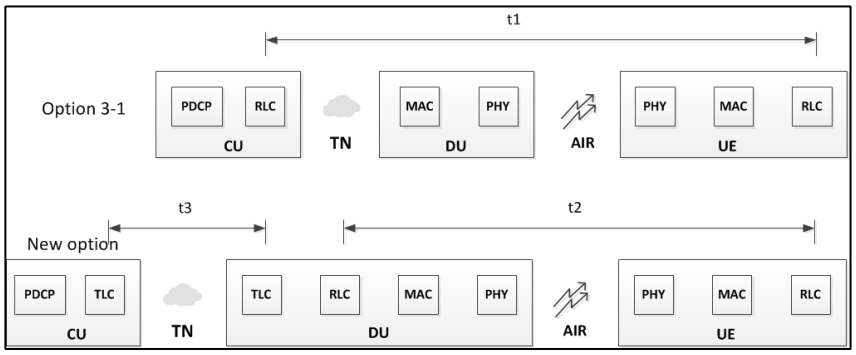


Figure 2-8 Delay in high layer split with and w/o TLC

The following relations apply:

Delay1 = 2*Ndrp (TN)*t1 + 2*Ndrp(AIR)*t1

Delay2 = 2*Ndrp (TN)*t3 + 2*Ndrp(AIR)*t2

Delay 1 – Delay 2 = 2*Ndrp (TN)*(t1-t3) + 2*Ndrp(AIR)*(t1-t2) > 0

2.3.3.12 Conclusions

The use of TLC under C3 control has the following benefits:

- Cellular network becomes aware of the operation of the TN based on the reported
  - Dropped packets (due to congestion)
  - Delay statistics
- Cellular network can reduce the traffic such to match the TN traffic capabilities
- Cellular network can execute handovers for selected traffic such to meet traffic QoS and slicing
- TLC introduction enables the reduction of the latency of the user data delivery.

2.3.4 Control framework for enforcing policy rules over CU-DU interface

2.3.4.1 General case

CU-DU split creates the problem of reproducing the buffer structure for user data at the DL receiving side of the CU-DU (F1) interface. For understanding, the relevant part of Figure 6-1 from TS 36.300: Layer 2 Structure for DL from TS36.300 [12] is reproduced below:

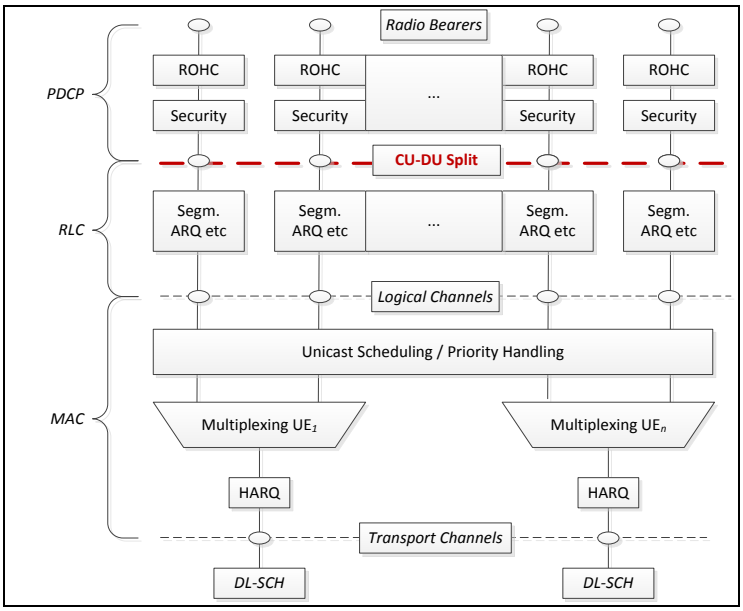


Figure 2-9 (Fig. 6.1 in TS 36.300:Buffer chain)

So when transmitting the user data of the F1 interface it is also needed to specify to which buffer the data belongs to.

The flow control used in X2-UP protocol creates a bearer for each UE and QCI. The flow control, including in Dual Connectivity the eventual error correction by retransmission, is based on PDCP PDU SN (Sequence Number) given for each UE and the QCI.

The 5G requirements in TR 38.913 [15] specify that "The target for connection density should be 1000000 device/km2 in urban environment".

Based on the core network architecture in TS 23.501 [10] for each UE could be several PDU Sessions and for each PDU Session could be several Radio Bearers. Additionally, for each Radio Bearer could be several QoS flows. If for each lowest hierarchy IP packet will be given a SN (Sequence number), it will result tens of millions of flows per km^2.

Handling such a huge number of SNs is not convenient for the implementation within the centralized gNB. Given that an UE can be served by multiple cells and by multiple DUs, may become problematic to find the relevant PDCP SNs for a DU in a short enough time.

2.3.5 User plane integration in centralized deployments

2.3.5.1 General case

A simpler approach is to define a Transport Block specific for F1 CU-DU interface. This TB will include the PDCP PDUs per QoS flow or per multiple QoS flows, eventually belonging to different PDU Sessions and UEs.

The size of this block should be chosen such to satisfy two criteria:

- Minimum lost PDUs in case of network congestion, which is accomplished with small PDUs;
- Maximum efficiency, which is accomplish with big TBs.

The structure of this TB shall be defined such to allow the selective retransmissions of PDU Blocks.

For the high layer split, the relevant DL PDU is produced by the PDCP layer and the UL PDU is the output of RLC layer.

The information carried out by each PDU block should contain not only the PDU, but also information which will allow placing the PDU content in the correct buffers at the receiver and also to further transmit the identifier if needed.

Such information includes:

- UE F1 identifier (or other suitable UE identifier such as UE X2 identifier)
- PDU Session identifier
- Network slice identifier
- Radio bearer (RB) identifier
- A QoS flow identifier, for example the 5QI.

A key point to clarify is how to use the error correction by re-transmission and how the gNB-CU will find the data to be retransmitted.

Based on TS23.501 [10] there shall be a flow for each UE and PDU Session and QoS flow marked with a 5QI (5G QoS Indicator) identifier.The user plane message coding is detailed in Annex A.

2.3.5.2 Short IP packets case

The short IP packets problem is described in Annex A.When only one PDU is transmitted over the F1 interface, a regular IP packet can be used for transporting it.

The TB may be formed either only from short PDUs or from a mix of long and short PDUs, for/from the same or different UEs, all using the same DSCP corresponding to the same or similar 5QI values. However each PDU Block in the TB should include its original 5QI.

Another grouping policy is based on enforcing (i.e. possibly different from the initial dropping probability) the same drop probability for all the PDU blocks in the TB, belonging to the same or different UEs. However the original 5QI should be transmitted in the PDU blocks.

In another solution, the PDUs with same drop probabilities may be grouped in the TB, instead of grouping based on 5QI. The SCTP value in the IP header transporting the PDU Blocks shall reflect the common PHB, while the bits for per-hop-behavior should reflect the most stringent 5QI behavior in the group.

The general principle of TB formation from different PDU blocks should be understood and be used in a flexible mode with regards to PDU size, PDU session ID, Network Slice ID, QoS indication such as CQI, 5QI, DSCP.

It should also be understood that a cellular entity can extract PDUs from a TB received from another cellular entity and regroup them based on a different criteria, which providing the additional information in a TB or PDU block in accordance with the criteria for regrouping.

Special applications

A special application is IoT, in which a Server polls a high number of devices for obtaining up-to-date reports. From the experience of such Server providers, there are a high number of dropped packets. These applications are not delay sensitive, such that remote devices send in general the reports as Best Effort, implying high drop probability.

Based on the fact that the reports are in general short and based on PDU session detection, the DU or a forwarding function or entity in the wireless access or core, could encapsulate the IP packets going in up-link and having all the same destination address based on the principle of PDU blocks,

where the actual PDU represents the initial IP packet and possibly the SSN and the checksum. The SCTP of the resulting IP packet shall use a coding for low dropping probability.

Connection to an Application Server

In the cellular network there are entities which in downlink forward the traffic to UEs and in uplink forward the user traffic to application Servers.

In case of IoT, the IP traffic consists of short packets and it makes sense to drop the overhead of the IP address and to increase their security. To this end, the pair entities can apply a protocol for replacing the IP header with a replacement identifier, like in Robust Header Compression, and creating PDU blocks having as content the replacement identifier, the IP message payload and eventually a sequence number and/or a check-sum which can be also a CRC. A check-sum can be added to the entire resulting IP packet or only to the payload of the resulting IP packet.

The cellular pair entity for data transfer can be an access gateway or function for accessing the Internet in the core network, a gateway or function in the core network forwarding and routing the traffic to a base station, or a base station which is stand alone or is composed from Central Units and Distributed Units or an User Equipment.

A cellular entity or a control function in the cellular network should communicate with the pair application server for establishing the parameters of the Transport Block, for example the size of the TB, the number of separate payloads (PDUs) to be included in TB, the size of each PDU block, whether a sequence number should be added to each separate payload, whether a check-sum should be added.

The drop probability in the DSCP field of the final IP header should also be established.

The header of the TB should have flags indicating one or more from the list below:

- Entire header of the original IP packet is encapsulated in the PDU block
- Only the UE IP address or an UE identifier is encapsulated
- A compressed header is encapsulated
- A sequence number is added for the entire TB
- A sequence number is added for each PDU block
- A check-sum is added for the entire packet
- A check-sum is added per PDU block
- An S-NSSAI is included in the PDU block.

The cellular network can define a PDU Session corresponding to the Server IP Address and keep the mapping between the UE IP Address and the compressed header for each traffic direction (uplink or downlink).

The cellular network should also keep the relation between the UE identity and the UE compressed header.

The TB can be further encapsulated in the GTP tunnel as used in the cellular network.

The UE may act as concentrator of traffic from devices in a given area.

RLC protocol enhancement

The existing approach is that over the air interface, after eventually applying FEC and HARQ error correction, IF a PHY transmit block is still detected as having errors, the entire block is discarded and in RLC Acknowledged mode the entire block can be re-transmitted.

A further refinement could be the use of SSN and check-sum (including CRC) for each SDU in the Transmit Block to be considered also by the PHY layer. This will allow determining which RLC SDU was correctly received and which one has errors.

A further enhancement for the short packets could be to apply a check-sum to a group of RLC SDUs, formed at the RLC layer, or to a group of PDCP PDUs, as described above.

Mobility

In case of UE mobility one or more PDU sessions can be transferred from one DU to another DU.

The packet builder function of the F1 interface shall be informed by C3 , about the modification of UE and PDU function assignment for each DU and build the F1 Transmit Blocks accordingly.

2.3.6 Monitoring of user plane operation

The Reports detailed in relation with TLC operation can also be used in this case.

2.3.7 Control procedures for user plane operation

Based on the proposed transport approach, for the case of high layer split, can be created a transport block for a specific 5QI, the transport block including PDCP PDUs from one or more UEs, preferably all of them belonging to the same 5QI.

The 5QI, formed from both standardized QoS profiles and non-standardized QoS profiles indicate different reliability requirements and time-budget requirements for each flow. The reliability requirements shall translate in a different retransmission policy of user PDU over the F1-D interface, while the time-budget should translate at least in a priority for transmission and at best in a suitable scheduling.

Given the full flexibility of the proposal, the traffic over the F1 interface **can be scheduled** to match the QoS and network slice requirements for a multitude of UEs and with the lowest overhead. For RAN processing the Network Slice is the parameter giving high level directions for polices related to PDU Session and QoS handling, while all the other elements are on a secondary plane. The S-NSSAI ID (Single-Network Selection Assistance Information) may also play a role in the actual implementation of 5CI requirements. It is recommended that the scheduling of transmissions over F1-U will be controlled by C3.

3 Summary of metrics to generate COHERENT network graph

3.1 Introduction

The COHERENT network information model [4] [3] is designed to support local short-term and long-term control at RTC and C3 levels, respectively. The concept of the network graph is a data structure which is populated by network information functions (NIFs) [3], operating at different levels of decentralization and time-scales. The possibility to deploy distributed and locally operating NIF instances enables smart aggregation of monitoring information at high degree of granularity and precision, as well as flexibility to combine data sources from various locations and levels into local and centralized network views. Storing monitoring information and network graphs effectively is enabled by the use of distributed databases capable of graph operations, such as NEO4J D6.1 [8]. Updates of network information are expected to be performed both synchronously and asynchronously as needed by pushing and pulling information from relevant NIF instances.

Using network graphs for representing different aspects of the observed network state and performance, allows for employing advanced graph algorithms, but also for effectively storing the relations between states and entities represented by nodes and edges for making effective coordination decisions (locally and centrally). Online and in a running system, a logically centralized C3 controller would, for example, operate on combined network views, to create high-level performance and resource management policies. At local levels, such C3 policies would be enforced by RTCs operating on local network views. However, the network graphs can also be used for offline analysis and simulation of advanced research concepts on, for instance, steering traffic, device-to-device communication, multi-connectivity in mmWave networks, and traffic transmitted between users or from/to network-user, e.g. to proactively identify network behavior patterns and employ resource management policies accordingly.

WP3 contributions address methods combining existing metrics into complex information and abstractions enabling and supporting novel controller applications. The contributions of WP3 offer various levels of abstractions based on composite metrics and probabilistic models, built from existing metrics representing e.g. signal quality and resource availability.

3.2 Proposed abstractions

The overall objective of this section is to highlight commonly used abstractions that in particular are RAT-agnostic and hence useful for coordinated RAN control involving heterogeneous RATs. The definition of an abstraction in the context of COHERENT and WP3 is based on understanding the relations between single metrics, composite metrics and how they relate to RAT-agnostic and RAT-specific representations:

- An abstraction can be created from single metrics (such as measured delay or standard metrics such as RSRP/RSQP) or from (nested) composite metrics (e.g. SINR).
- An abstraction can be RAT-agnostic or RAT-specific
- A RAT-specific abstraction exposes information built on RAT-specific metrics and composites and are used by RAT-specific control functions.
- A RAT-agnostic abstraction is built on RAT-specific metrics transformed into information generic to several RATs, and can be used by RAT-agnostic control functions.

Note that the input and output from a NIF instance can be a single or a composite metric – depending on the level of abstraction and its complexity, an abstraction can be the result of chained or hierarchical NIF processes.

Overall, proposed or used abstractions represent instantaneous or predicted aspects of the network state and performance, covering mainly three categories of RAT-agnostic abstractions:

- Link quality and performance (e.g. attainable throughput, signal strength, interference)
- Topological properties (e.g. distances between client and server nodes)
- Physical infrastructure parameters and resource capacity (e.g. available radio resources or estimated load of serving nodes, equipment parameters, such as, interval of transmission powers, computational capacities, etc.).

These are general RAT-agnostic abstraction categories needed for high-level coordination and 5G RAN-control, which can be built from RAT-specific metrics. The network information concept of COHERENT allows for flexible compositions of various abstractions relevant to specific controller applications (section 4).

Table 3-1, lists generic RAT-agnostic abstractions (created from existing metrics and measurements) that can be used by various controller applications. Although contributions in section 4 have been evaluated in a certain context, the common denominator of most of these abstractions is the generality of the information exposed and applicability to any radio access technology or control application. Derived abstractions can be viewed as constituting a set of relevant information items needed for coordinated control of heterogeneous 5G RAN.

Table 3-1 RAT-agnostic abstractions used in COHERENT

Category	Abstraction	Section
Link quality and connectivity	Signal-to-noise ratio (SINR/SNR)	4.1.1, 4.1.2, 4.2.1, 4.2.2, 4.3.2, 4.3.4, 4.4.2, 4.4.3, 4.4.4
	Interference	4.3.3, 4.3.5, 4.4.4
	Throughput	4.1.1, 4.2.2, 4.3.1, 4.3.2, 4.4.1
	Latency	4.3.2
Topological properties	Inter-node distance	4.3.3
	Coverage radius	4.3.3, 4.4.2
	Node distribution density	4.3.3
	Relative speed between nodes	4.3.4
Physical infrastructure parameters and resource capacity	Data rate (bit/s)	4.1.2
	Bandwidth (Hz)	4.1.1, 4.1.2, 4.2.1
	Spectral efficiency (bit/s/Hz)	4.1.2, 4.3.5
	Number of antennas	4.3.4

	Transmission power	4.4.2
	Path loss	4.4.2
	Energy consumption	4.4.1

In addition to the abstractions used in Table 3-1, we introduce high-level abstractions related to resource usage and link quality and performance in Table 3-2. Although the notion of these high-level abstractions can be found in the literature in different contexts, novel expressions have been developed within the scope of new COHERENT control applications and techniques. Here, "high-level" refers to a function of abstracted input (e.g. such as a distribution). These are all (except the "Received power" abstraction) created using probabilistic approaches. Probabilities, expressing the relative performance or resource usage of individual network entities, effectively provide a normalized RAT-agnostic information model applicable in highly heterogeneous network environments. Additionally, probabilistic representation of the network state is useful, adding an additional abstraction level allowing for specifying controller behaviors and policies in an expressive manner through probabilistic requirements.

Table 3-2 Novel expressions of high-level abstractions produced in COHERENT

Category	Abstraction	Section
Link quality and connectivity	Estimated received power	4.3.4
	Probability of coverage	4.4.2
	Attainable throughput	4.4.1
	Weighted achievable throughput for sidelink communications	4.3.1
Physical infrastructure parameters and resource capacity	Resource depletion probability	4.1.1
	Risk of overload	4.1.3

Details on approaches employed for creating variants of listed abstractions and control applications are presented in section 4. In addition to the abstracted metrics in Table 3-1and Table 3-2, section 4.2.3 also proposes API abstractions for e.g. transmission policies.

4 COHERENT applications of abstraction and control methodologies

The general direction of WP3 involves the generation of network graph abstractions and control applications using these abstractions. In this section, a wide variety of applications and use-cases is presented to evaluate the applicability of network graph concept with COHERENT architecture for PHY/MAC abstraction. The details of 16 technical contributions are presented, grouped in four categories of control applications: load balancing, multi-connectivity, coverage extension and link performance. For the convenience of the reader, we briefly summarize the direction of each contribution as follows:

Load balancing

Sec. 4.1.1 with focus on load balancing in dynamic multitier heterogeneous networks presents a probabilistic abstraction describing the stochastic behavior of the resource consumption of BSs of different tiers in a heterogeneous cellular network. The goal is to provision the probability that users experience connection blockings due to resource depletions in the BSs.

Sec. 4.1.2 load balancing using traffic steering in HetNet is a novel approach showing that centralized load balancing based on network graph view improves the distribution of network load, and it is able to reduce the number of unsatisfied UEs and to increase network spectral efficiency.

Sec. 4.1.3 presents probabilistic load balancing and employs overload risk estimates which can be used for exposing this aspect of the network state in an abstract way, while efficiently reducing the occurrence of high loads and limits the balancing effort and cost compared to similar approaches.

Multi-connectivity and multi-casting

Sec. 4.2.1 presents multi connectivity in LTE, which exploits the hierarchical control architecture for both centralized and real-time rate allocation control, showing benefits in both network and user perspectives in terms of QoS guarantee and satisfaction ratio.

Sec. 4.2.2 centered around multi-connectivity in mmWave networks and LTE, aims at a future 5G system addressing network management problems related to mmWave & sub-6GHz multi-connectivity – by dividing the optimization problem into two levels, solving cell association problems centrally and resource allocation problems at RTC), the computation overhead for network optimization is shown to be reduced while satisfying 5G user experience requirements.

Sec. 4.2.3 orients on multicasting in WiFi with focus on a novel multicast rate adaptation and mobility management scheme for 802.11–based WLANs - the proposed scheme uses an SDN approach where the global network view available at the C3 is exploited for coordinating the operations of different APs.

D2D communication and coverage extension

Sec. 4.3.1 gathers the results on studies on in-coverage and out-of-coverage D2D communications and relaying topic mainly related to the concept of moving cell or to in-coverage relaying for increasing the performance of cell-edge users – a flexible weighted metric of throughput is presented applicable to e.g., predict the maximum number of users successfully accessing a medium or to configure D2D communication parameters relative to current access demands.

Sec. 4.3.2 focuses on coverage extension through D2D for LTE and evolutions, and shows how the COHERENT architecture could be applicable to UE-to-Network relays, in particular for achieving coverage extension through D2D communications.

Sec. 4.3.3 deals with in-coverage D2D communications, using two-hop D2D relaying to enhance the DL end-to-end throughput in a multi-cell context - an analytical model is provided of the aggregate co-channel interference when D2D relaying, with a set of parameters including a minimum relaying distance and a relaying probability which enable the control of D2D activity.

Sec. 4.3.4 performs a mobility study for a high speed platform in the context of air-to-ground communications, using power measurement strategies and network graph information for dealing with high Doppler shifts, relevant for e.g. effective handovers.

Sec. 4.3.5 further studies the moving cell concept in the context where existing LTE features are combined to build an enhanced eNB (e2NB), which can be used as a node in a wireless mesh network without infrastructure - centralized and local scheduling at different time-scales enables effective resource management while fulfilling QoS requirements for elastic flows.

Link performance

Sec. 4.4.1 proposes a probabilistic approach towards building RAT-agnostic network graphs based on attainable throughput for WiFi and LTE. The probabilistic metric supports controller applications making decisions on performance management and SLA enforcement for individual UE connections.

Sec. 4.4.2 describes a DAS system model as well as the probability of coverage metric derived for analyzing different DAS scenarios i.e., varied number of RRHs in a cell and targeted SINR at different UE positions. The probability of coverage is used to ensure that the required throughput at the UE location is achieved statistically based on the channel model.

Sec. 4.4.3 shows how the abstraction and network graphs can be used for new technologies, such as massive MIMO within the COHERENT architecture and for heterogeneous mobile networks. Exploiting this new technology together with a centralized coordinator allows for better performance for delay sensitive applications mentioned in 5G recommendations.

Sec. 4.4.4 proposes an inter-RAT scheduling algorithm to share limited time-frequency resources among users utilizing various RATs, namely 2G and 4G in this case. The proposal can be of high importance in dense heterogeneous network utilizing limited frequency resources.

Sec. 4.4.5 proposes a logical description of the RAN by Logical RAN Entities, and describes the state of the RAN by a network graph, to solve disputes for resources in HetNets implementing ICIC at network-level. Results show the feasibility of performing network-level coordination with the proposed approach, and positive gains for users in the lower percentiles, without loss in system average user rate.

4.1 Load balancing

4.1.1 Load balancing in dynamic multitier heterogeneous networks

4.1.1.1 Motivation and goals

A rather new feature of next generation cellular networks is that they are becoming highly heterogeneous, dynamic, and unstructured. As a result, a regular and static grid cannot anymore represent the heterogeneous topology with different cell sizes, as it was common in conventional cellular networks using only macro base stations. In addition to randomness in position of users and base stations, the effects of randomness of user traffic flows and resource availability must be well understood by sophisticated resource coordination approaches. The work reported in this section as well as related work in D3.1 [4] supports the applications such as resource sharing among HetNets and coordination of rapidly deployable mesh networks defined in WP2.

The main goal of this work is to develop traffic steering schemes with an emphasis on load balancing between different tiers of dynamic and unstructured cellular heterogeneous network (HetNet). The work is a follow-up of the work reported in deliverable D3.1 [4] which focused solely on probabilistic energy graphs for centralized load balancing where heterogeneity was primarily considered as the energy access characteristics of the base stations (BSs). Some related

work can be found from [23] and the references therein. A more comprehensive literature review is provided in D3.1[4]. More generally, the heterogeneity can be understood as the distinctive characteristics of cellular base stations of different tiers in terms of spatial density, transmit power, inter-tier interference, usage of traffic-dependent resource blocks, and accessible amount of source (e.g. solar or wind) energy in BSs. Furthermore, the heterogeneity from using non-cellular radio access technology, such as WiFi, can be considered as a potential extension of the main work.

4.1.1.2 Requirements

The target control approach includes both the top-down process (i.e. control action from C3 to BSs) and the bottom-up process (i.e., network status information from BSs to C3). The top-down part involves the traffic steering decision-making algorithms and identification of required parameters that need to be gathered by the network graph. Load balancing is understood as a method to avoid or minimize resource depletions in the BSs where resource type of interest can be, e.g., physical resource block (PRB) or energy consumption in the selected energy unit. This is achieved by adapting the number of associated users given the heterogeneous resource status of involved BSs. At the same time, it is desired that user-specific goals, such as user throughput, are taken into account. The type of a BS, e.g., on-grid macro base station (MBS) or off-grid small base station (SBS), affects the importance of the resource type in the load balancing process. Instead of typical static deployment of BS locations and user traffic models, we focus on dynamic HetNet topology and traffic models. This calls for probabilistic approaches to abstract the network state information that is exchanged between non-real-time C3 and local real-time controllers (RTCs) in order to create sophisticated network graph for C3 decision making.

4.1.1.3 Parameters and applied abstractions approaches

The main low-level measurements of interests, which build up the target network graph for C3, are the PRB and energy consumption to define the dynamic resource usage at different types of BSs as well as the signal-to-interference ratio (SINR) to indicate the achievable throughput. Specifically, we apply probabilistic approaches to characterize resource usage on PRB and energy. The main objective of applied probabilistic metrics is both to compress information and to include finite-horizon prediction capability of the consequences of the C3 decision-making process. The list of parameters of interest is provided in Table 4-1. In this work, we focus on downlink (DL) direction only.

The main idea of the applied abstraction approach is to use a probabilistic measure to evaluate the potential resource depletion in the near future for the target finite-horizon window in time units. Knowing such a measure may help to evaluate the risk of performing certain control action or it may help to evaluate the need to change the control approach, e.g., from a distributed to centralized mode that can more effectively address the resourcing problem by offloading traffic to different tiers with global network status information.

Generalizing the energy graph concept from D3.1 [4] to a more general resource graph concept, the number of resource units (e.g. resource blocks or energy units) in the *j*th BS at the *t* can be represented as

$$q_j(t) = \min\{\max\{b_j + x_j(t) - y_j(t), 0\}, B_j\} \qquad (4\text{-}1)$$

where $x_j(t)$ is the arrival resource unit process, $y_j(t)$ is the departing resource unit process, b_j is the initial number of resource units, and B_j is the maximum amount of resource units available to the BS. A resource depletion event is declared if $q_j(t)$ becomes zero. Provided that B_j is large enough, its effect on resource depletion probability becomes negligible. Given the resource buffer status at $t = t_0$, the main objective is to be able to predict the depletions before time $t = t_0 + S$ where S is the target prediction horizon in time units.

The resource depletion probability (RDP) is defined as the probability that $q_j(t)$ becomes zero for given mean and variance of resource unit inter-arrival and inter-departure times. To simplify the calculation in a closed form for online processing, it is shown in [22] that the RDP with finite prediction horizon *S* can be well approximated by

$$P_j = \int_0^S \frac{b_j}{\sqrt{2\pi\alpha_j t^3}} \exp\left[-\frac{(\beta_j t + b_j)^2}{2\alpha_j t}\right] dt \approx \omega_j \exp\left(-\frac{2\beta_j b_j + b_j^2}{2\alpha_j S}\right) \qquad (4\text{-}2)$$

where ω_j, α_j, and β_j depend on the measured moments of $q_j(t)$ and are given in [22]. In essence, the RDP conveniently hides (abstracts) the current resource status along with its expected statistical behavior of the incoming and outgoing resource units within the selected time window into a single parameter. In Table 4-1, SINR and throughput are calculated based on standard methods using received signal powers from desired and interfering base stations and Shannon capacity formula, respectively.

Table 4-1 List of parameters of interest

Parameters	Description	Method	Dependence	Deployment
RDP	Finite-horizon resource depletion probability of BSs	Abstracted	PRB, ECU	RTC
SINR	Received mean signal-to-interference ratio at UEs	Abstracted	RSRP	UE
TP	Mean throughput of UEs	Abstracted	SINR	UE/RTC
PRB	Number of available physical resource blocks at BSs	Measured	-	RTC
ECU	Number of available energy consumption units at BSs	Measured	-	RTC
RSRP	Received power at UEs from BSs	Measured	-	UE
UAD	User association decision vector	Controlled	RDP, TP	RTC/C3

4.1.1.4 Applied load balancing approaches and use cases of network graphs

Load balancing scheme is often based on averaged information of the status of BSs. A smarter load balancing scheme uses the historical status information to predict the consequences of the decisions for the performance in the near future. The low-level instantaneous measurements are therefore transferred into probabilistic measurements that can be delivered to C3 in a slower timescale to reduce the control overhead in using a centralized control approach. This information is then used in the traffic steering control to balance the load so that the probability that resources (i.e. PRB or energy) of the BS are depleted is minimized. Alternatively, the probabilistic measurements can be used to trigger C3 proactively and on-demand basis so that user association is made in a distributive manner when applicable to reduce control signaling. In general, the centralized load balancing can benefit from global knowledge of network status information provided by network graphs to make optimal traffic steering decisions.

A block diagram of the target system model is shown in Figure 4-1. It consists of three levels including radio access level involving local measurements, RTC level involving local abstraction process, and C3 decision level involving global network graph construction. The bottom-up process starts from the radio access level, delivering instantaneous network measurements to RTC level which further abstract the information for the C3 level. The top-down process sets the user association decision vector to balance the load which is conveyed to RTC level after which the decisions are executed.

In the following, we consider two use case studies where Case 1 is related to user association to off-grid multi-hop small BSs by using probabilistic energy graphs of BSs in C3 and Case 2 is related to user association to multitier BSs by using throughput graphs of UEs along with proactive triggering of C3 with RDP of physical resource blocks.

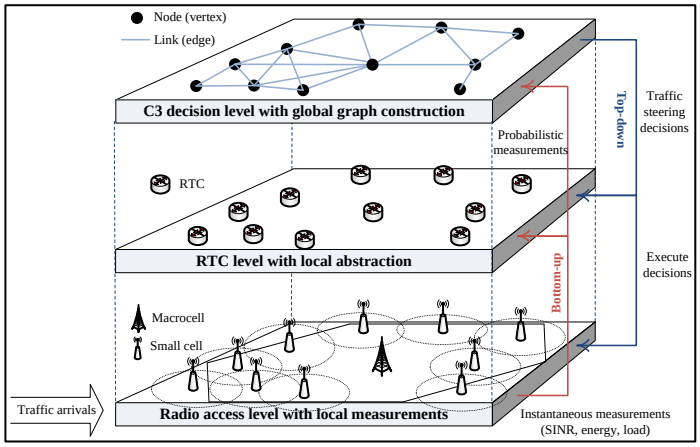


Figure 4-1 Block diagram of target system model

In the first study case, we focused on C3-based load balancing for avoidance of energy depletion in off-grid small cell flexibly deployed mesh networks. This part of the work was reported in D3.1 [4]. In this second study case, we focus on load balancing for UEs throughput outage minimization in multitier cellular networks with the help of probabilistic PRB depletion analysis. The main objective is to evaluate the finite-horizon RDP for a typical BS in different tiers of the network as the traffic intensity increases. This information can be used to trigger centralized user association approach. To model the unstructured and dynamically changing network load, we use the stochastic geometry framework where both the BSs and UEs are randomly located as well as queueing theoretic framework where UEs arrive and depart the network at random time instants.

The applied three-stage control approach is as follows. To reduce the control information it is desirable that whenever applicable, UEs can associate to BSs in a distributed manner using only the local received signal power information between UEs and BSs. We apply the standard biased received power for the distributed user association approach where a UE is connected to the tier that satisfies $\text{argmax}_k\, \theta_k \mathcal{P}_{\mathrm{r},k}$ where θ_k is the selected association bias of tier k and $\mathcal{P}_{\mathrm{r},k}$ is the mean received signal power from the best BS of tier k. As the traffic density becomes higher, the risk increases that BSs, which associate with UEs distributedly, deplete their resources. Therefore, each BS calculates the RDP to indicate whether a more sophisticated user association performed by C3 is needed. The RDP is evaluated over finite time horizon which is set to be the expected time delay from requesting the control decisions from C3 to time instant where the control decision is in action. If the RDP of a given tier exceeds the target risk threshold, the bias parameters of user association metrics can be updated by C3 or alternatively C3 can perform the decision making process completely. For the latter approach, C3 solves the association decision

vector $x \in (0,1)$ by maximizing $\max_x \sum_i \sum_j x_{ij} \log c_{ij} - \sum_j L_j(x) \log L_j(x)$ where c_{ij} is the throughput of UE i connected to BS j and $L_j(x)$ is the load of BS j which depends on the decision vector, cf. the approaches in (Andrews 2014) and the references therein. The throughput can be measured using the well-known Shannon rate-power mapping with RAT-specific parameter scaling factors to incorporate any non-idealities of practical modulation and estimation methods. There are several approximate approaches to reduce the complexity of this maximization problem. For instance, the integer decision vector can be relaxed to take any value real value within [0, 1], allowing to use simpler optimization tools. The throughput can be measured for a selected subset of UEs and BSs instead of calculating it for every UE or BS separately.

4.1.1.5 Numerical results

In this subsection, selected numerical results are provided for Case 2 using a simulation platform based on Matlab. Numerical evaluation of Case 1 can be found from D3.1 [4]. In generating the results, the following assumptions are made. We consider 2-tier (i.e., MBSs and SBSs) cellular network where the BSs of each tier (tier represents a class of BS having unique features) are located randomly according to spatial Poisson point process with *k*th-tier spatial density λ_k (λ_1= 1 BS/km^2, λ_2= 10 BS/km^2) and transmit power $\mathcal{P}_k$ ($\mathcal{P}_1$= 53 dBm, $\mathcal{P}_2$= 33 dBm). UEs are located randomly both in space and time forming a spatiotemporal Poisson point process with a given spatiotemporal density defined as mean number of UEs per space (km^2) and time (sec) unit. The standard distance-dependent path loss model with path loss exponent of three is assumed. The center frequency range of interest is between 2 and 4 GHz where different tiers are allocated separate frequency ranges but frequency reuse factor of the BSs in the same tier is set to one. The link fading is assumed to be non-frequency-selective and to follow the Rayleigh distribution. The non-delayed access scenario is assumed where the control unit denies the UE access if there are no resources to be immediately allocated. Admitted UEs stay in the system for an exponentially distributed time duration with rate μ which depends on the file size and cell capacity.

In Figure 4-2 (left-hand side) BS-level, i.e. single BS isolated from the network with ten remaining resource blocks, analytical and simulated RDPs are illustrated as function of mean number of arriving UEs in a given spatiotemporal range for different prediction time horizons which represent in this case the expected response time of C3. The RDP increases with prediction time horizon. It is also seen that the proposed analytical abstraction approach can follow the simulation results quite closely while providing a simple way to calculate the RDP. In Figure 4-2 (right-hand side) we depict the tier-level RDP when using actual distributed user association approach where tier 1 becomes congested and tier 2 is not. The tier-level RDP represents the averaged RDP of a typical BS in a given tier and provides interesting insight on the risk of depleting energy in different tiers for given tier-specific parameters. Moreover, it can be used to predict the moment when it would be advisable to change the distributed user association mode into centralized mode with the hope of leading to an improved load balancing state in the network. In Figure 4-3, we evaluate the UE throughput outage probability with threshold of 100 Mbit/s including all tiers using the distributed and centralized modes for the user association. The resource outages are reflected either by non-blocking mode by reducing the bandwidth of UEs to fit all UEs in associated cells or by blocking those users arriving after all resources are in use. It is seen in Figure 4-3 that when blocking is used, the centralized control has the highest gain compared to distributed control. This motivates to use resource outage probability as a good indicator for C3 triggering.

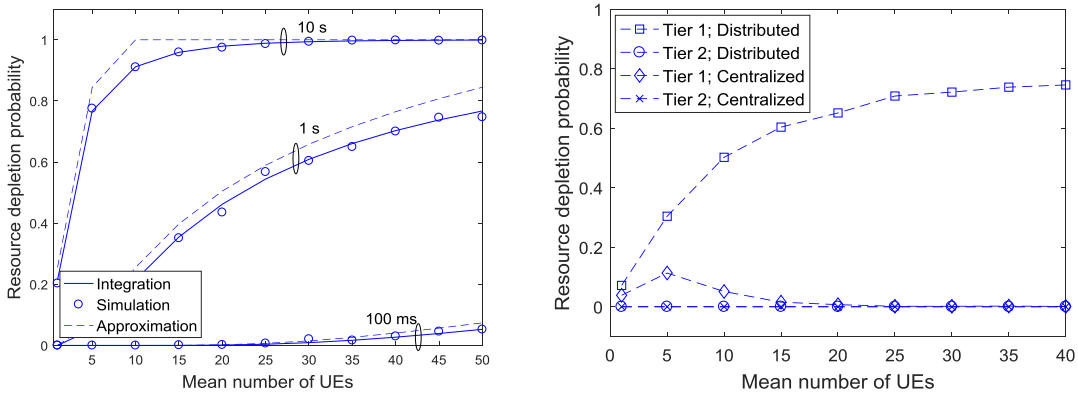


Figure 4-2 BS-level and tier-level analytical and simulated RDP for different prediction time horizons (i.e., expected response time of C3) as function of mean number of UEs

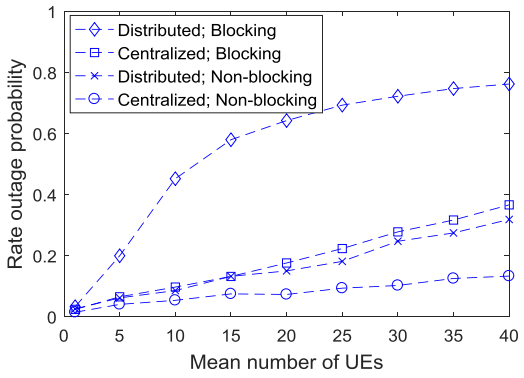


Figure 4-3 Network-level simulated UE rate outage probability (threshold 100 Mbit/s) for distributed and centralized user association as function of mean number of UEs

4.1.1.6 Conclusions

In this study, load balancing via traffic steering in dynamic multitier heterogeneous networks is investigated. The proposed probabilistic abstraction approach connects several parameters into single parameter, describing the stochastic behavior of the resource consumption of BSs of different tiers in the cellular network. The goal is to provision the probability that users experience connection blockings due to resource depletions in the BSs. Two case studies are represented for which the abstraction approach is applied. In the first case, we address abstraction of energy status as energy graphs of small base stations harnessed with specific energy charging capability from the surrounding environment. In the second case, the probabilistic approach is applied to predict the BS-level and tier-level resource depletions to ask help from C3 for improving the user association process. It is concluded that the use of C3 may help to balance the load leading to balance in performance of different tiers. However, with lower traffic loads, distributed approach can lead to satisfactory results, suggesting using predictive on-demand approaches to trigger the use of C3 on-demand basis to reduce control signaling as an alternative to use of C3 periodically.

4.1.2 Load balancing using traffic steering in HetNet

4.1.2.1 Motivation and goals

In an operational wireless heterogeneous network, the traffic is not evenly distributed. Some areas or frequency bands may be congested while others are underutilized. With load balancing, a target is set to achieve a balanced network load by steering traffic from the highest loaded cells to cells with spare capacity. When one RAN node becomes congested, the load shall be balanced by handing over its associated end user devices to other neighboring RAN nodes. The goal is to improve spectral efficiency of the network and offer higher throughputs for users.

4.1.2.2 Requirements

a. Abstraction of available capacity per RAN node

To balance load between RAN nodes, it is crucial to define an abstract estimate of the available capacity per RAN node so that the load of RAN nodes can be properly compared in the load-balancing algorithm.

b. Centralized approach

A centralized load balancing is considered as more suitable for following reasons:

- Distributed load balancing mechanisms are only standardized for specific RATs (e.g. for X2-based load balancing in LTE; Mobility Load Balancing SON function), while for legacy 2G, 3G and WiFi systems it is not the case. As COHERENT project targets load-balancing in multi-RAT heterogeneous network, there is a requirement to use centralized load balancing.
- A centralized load balancing approach can realize more consistent optimization applied across RAN nodes from multiple vendors.
- COHERENT project introduces the concept of COHERENT coordinator so it looks quite logical to use this architecture component for load balancing tasks.
- A centralized load balancing approach/algorithm can use the additional Key Performance Indicator (KPI) information available only at the management plane and not at the RAN nodes.

c. Abstraction of end device measurements

The channel measurements are technology-specific: for example, LTE UE measurements include RSRP and RSRQ, 3G measurements include RSCP, EcNo, and WiFi measurements include RSSI. Technology-specific measurements should be abstracted to be comparable.

d. Dynamic adaptation

Load balancing should be able to dynamically react to the network changes. This imposes a requirement for the fast load balancing algorithms with reaction time comparable with load balancing in distributed architecture (performed locally in RAN node).

4.1.2.3 Generation and usage of Network graphs

Each node at the graph represents RAN node (abstraction of WiFi AP and LTE eNodeBs) and the user device (UE and WiFi-capable device). The edge between the end device and the RAN node represents a link quality between the RAN node and the user device if the RAN node can be considered as a candidate for serving cell. If the link quality is below a threshold (that is technology-specific) then the RAN node cannot be considered as a load-balancing candidate and the edge is not present. The selected edges are marked as serving links and represent current association(s) of user device with RAN node(s). RAN node is attributed with Available Node

Capacity (ANC). This graph representation is technology-agnostic and allows to perform load-balancing in technology-agnostic manner.

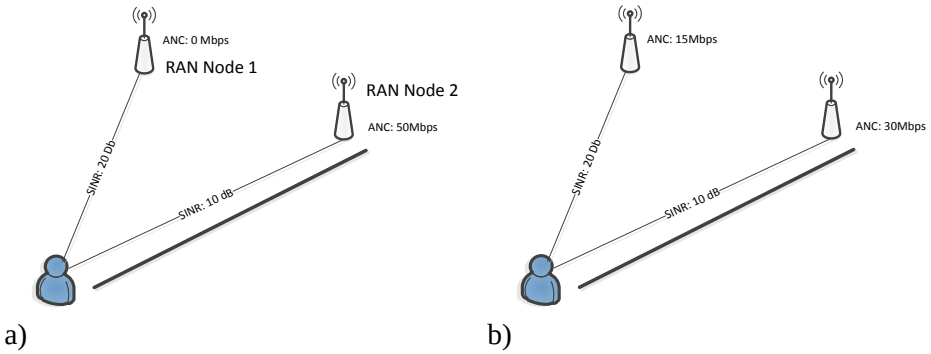


Figure 4-4 Graph representation for load-balancing

For example, based on the graph representation in Figure 4-4, C3 can decide to change association of user device from RAN Node 1 to RAN Node 2 as RAN Node 2 is underutilized (ANC:50 Mbps) while RAN Node 1 is fully utilized (ANC: 0 Mbps). Signal quality between nodes (SINR) is also respected to select the best candidate for load-balancing.

4.1.2.4 Abstracted parameters

Available Node Capacity = Available bandwidth x Spectral efficiency

Available bandwidth *(in Hz)* indicates the amount of frequency resources that are available at a RAN node. Available bandwidth is also impacted by backhaul load, control channel capacity, CPU load, and current QoS satisfaction requirements.

Spectral efficiency (in (bits/s)/Hz) is the average bit rate that can be transmitted over a given bandwidth. It is a measure of how efficiently frequency resources are utilized at RAN node.

Both *Available bandwidth* and *Spectral efficiency* are measured and averaged before reporting. Reporting and averaging is executed in time-scale of seconds. Available node capacity is defined for uplink and downlink: ***Available_Node_Capacity_UL*** and ***Available_Node_Capacity_DL.***

In terms of available node capacity, thresholds ***Available_Source_Node_Capacity_DL_Thesh and Available_Source_Node_Capacity_UL_Thesh*** are defined. When ***Available Node Capacity*** of a RAN Node falls below these thresholds, the coordinator shall try to offload RAN Node.

The parameter ***Max_Handover_Number_From_Node*** limits a number of load-based handovers from RAN Node at each decision round to avoid situation that too many end devices are handed over due to load-balancing and in the next round a RAN node become almost free.

The parameter ***Max_Handover_Number_To_Node*** limits a number of load-based handovers to RAN Node at each round to avoid situation that too many end devices are handed over due to load-balancing to RAN Node and in the next round a RAN node become likely congested.

When there are less than ***Min_Num_Of_End_Deivices,*** then no end devices will be handed over.

Table 4-2 Exposed parameters

Parameter	Description	Direction	Deploy-ment
Available_Node_Capacity_UL	Available capacity per RAN node for Uplink	UL	C3
Available_Node_Capacity_DL	Available capacity per RAN node for Downlink	DL	C3
Available_Source_Node_Capacity_DL_Thesh	Threshold of RAN Node Capacity RAN node for Downlink to consider offloading **from** RAN Node	DL	C3
Available_Source_Node_Capacity_UL_Thesh	Threshold of RAN Node Capacity RAN node for Uplink to consider offloading **from** RAN Node	UL	C3
Available_Target _Node_Capacity_UL_Thesh	Threshold of RAN Node Capacity RAN node for Uplink to consider offloading **to** this RAN Node	UL	C3
Available_Target _Node_Capacity_DL_Thesh	Threshold of RAN Node Capacity RAN node for Downlink to consider offloading **to** this RAN Node	DL	C3
Max_Handover_Number_From_Node	Maximum number of handovers that can performed from a RAN Node at each evaluation period (round)		C3
Max_Handover_Number_To_Node	Maximum number of handovers that can performed to a RAN Node at each evaluation period		C3
LB_Channel_quality	Channel quality to neighboring RANs measured from end divices		
Min_Num_Of_End_Devices	Minimum number of connected end devices to consider RAN Node for offloading	UL/DL	C3

Table 4-3 Controlled parameters

Parameter	Description	Direction	Deployment
Serving Cell	Serving RAN Node of end device	UL/DL	RTC, C3

4.1.2.5 Algorithms in C3

The interaction with C3 for technology-agnostic load balancing is shown at Figure 4-5. The technology-specific measurements reports from end devices (containing channel information of neighboring cells) and resource utilization of RAN nodes will be passed to abstraction module where they will be abstracted into SINR and Available Node Capacity for technology-agnostic graph construction at the C3. The C3 will evaluate load situation, find a good candidate for offloading/handover for selected end devices. Then, the C3 will request offloading/handover to balance network load with a Load balancing Request that is mapped to a technology specific handover command.

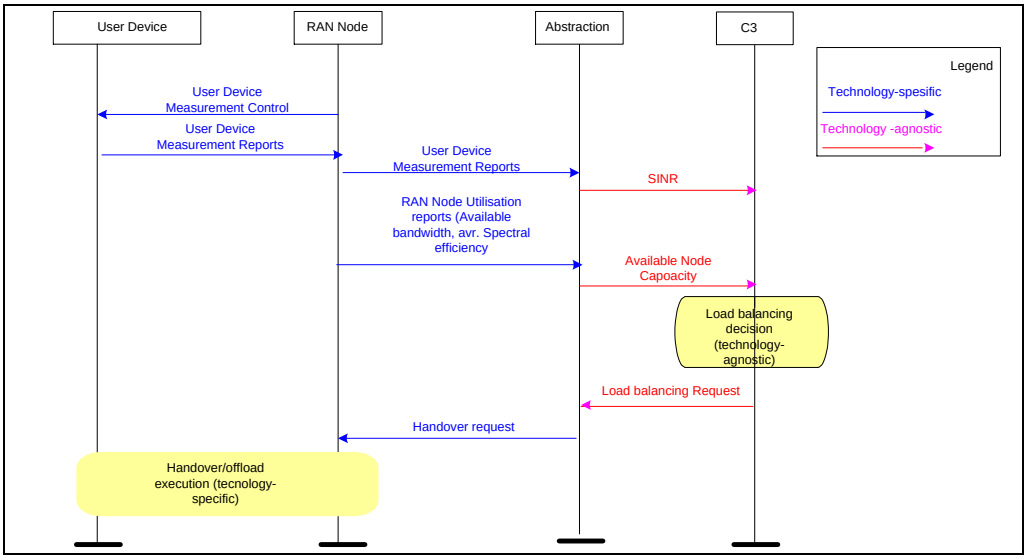


Figure 4-5 Block diagram representation

Load-balancing decision is done every T milliseconds. Algorithm steps (executed are for each RAN node) are:

1. The coordinator will check the Available Node Capacity for DL and UL: ***Available_Node_Capacity_DL*** and ***Available_Node_Capacity_UL.*** If they fall (at least one of them) below theresholds ***Available_Source_Node_Capacity_DL_Thesh and Available_Source_Node_Capacity_UL_Thesh*** correspondingly, proceed to the next step 2.
2. Check the number of UEs served by this RAN node. If number of connected end devices larger than ***Min_Num_Of_End_Deivices*** proceed to the next step 3.
3. While number of connected end devices at considered RAN node larger ***Min_Num_Of_End_Deivices*** while there are connected end devices UEs and neighboring cells satisfying stated below conditions, do the following:
   - Find a connected end device with the highest value of signal quality to neigbouring RAN Node's and larger than ***LB_Channel_quality***, for which ***Available_Node_Capacity_DL*** smaller than ***Available_Target_Node_Capacity_DL_Thesh*** and ***Available_Node_Capacity_UL*** smller than ***Available_Target_Node_Capacity_UL_Thesh***, and number of UEs that have been commanded to handover to the target RAN Node the current decision round smaller than ***Max_Handover_Number_To_Node***.
   - If UE is found in the previous step,

Send handover order for the end device to hand over it to the target cell

- Increment number of UEs that have been commanded to handover to the target RAN Node ***Max_Handover_Number_To_Node*** and ***Max_Handover_Number_From_Node***.
- If UE is found in the previous step,

Send handover order for the end device to hand over it to the target cell Increment number of UEs that have been commanded to handover to the target RAN Node ***Max_Handover_Number_To_Node*** and ***Max_Handover_Number_From_Node.***

4.1.2.6 Results

The benefits of load-balancing handover in LTE have been shown in many studies and deployments. Also, in a lab setup with 2 eNodeB and 2 UEs when 2 UEs are connected to one eNodeB based on radio condition at the beginning of the test, each UEs reaches about 23Mbps (MCS=24) at 20MHz bandwidth. After the enforced handover to neighboring eNodeB, UE could improve their throughput to 35Mbps in the target cell, and UE staying in the original cell could expand its throughput to 53 Mbps. An extension of the load-balancing scheme to heterogeneous Wi-Fi/LTE networks has been presented in [24].

4.1.2.7 Conclusion

In this work, it was verified that the traffic offloading decision can be shifted in technology agnostic manner to the coordinator. It was demonstrated that the standard LTE can be customised to provide the possibility to trigger handover not only from eNodeBs (as specified in the current LTE specification), but also from the coordinator. This study is considered a first prototype for virtualisation of handover (mobility) function moving it from eNodeB to the cloud network. The setup for verification has been implemented in WP6 based on the commercial eNodeB and the 5G-EmPOWER reference C3 implementation.

4.1.3 Load balancing using probabilistic models

4.1.3.1 Motivation and goals

The main goal is the development and integration of a self-organized and RAT-agnostic probabilistic load balancing mechanism into a logically centralized framework through scalable monitoring and a remotely configurable semi-autonomous rebalancing actuation mechanism. The features offered by the mechanism include:

- High-level control and monitoring of cell load via probabilistic risk abstractions.
- Scalable self-organization of load balancing based on levelling the risk of reaching a specified (threshold) load level.

The approach is aimed to support unified and programmable control of heterogeneous infrastructures by the means of the probabilistic risk abstraction and self-organization, while addressing the issue of scalability by operating in a distributed manner.

4.1.3.2 Requirements

The following requirements must be fulfilled for the mechanism to work for an arbitrary RAT

1. Momentary load measurements over a suitable time scale ts, (e.g. aggregated channel busy time, or consumed resource blocks).
2. A way for each node to identify a neighbourhood of other nodes (with which to interact in a distributed computation of target load values).

3. A mechanism for nodes to communicate load and target load values bilaterally.
4. A mechanism to bias the choice between alternative serving nodes for a terminal to connect to.

The mechanism is bottom-up in the sense that, once configured, the controller of each node operates autonomously, but also reports the metrics computed for its self-regulation to a RTC/C3 controller responsible for a larger number of nodes. This controller can, based on the reports from groups of node controllers influence the rebalancing actuations of individual node controllers through a well-defined control API.

4.1.3.3 Application of Network graphs

The proposed algorithm is designed to offload the high-level control layer by operating in a distributed manner, taking high-level requirements related to acceptable load from the C3. The local estimates of the overload risk can be exposed by the infrastructure in terms of network information functions (NIF) and collected (pulled or pushed) by requests via the north bound interfaces of the architecture as described in [3] and represented in terms of a network graph. The applications of network graphs representing the risk of overload can then be used for several applications over shorter and longer time periods, e.g. for monitoring or for learning load shift patterns and re-association rates for network planning, optimization and coverage adjustments.

The work here doesn't consider how this monitoring information is stored, aggregated or used to produce control parameter values p_i at the RTC/C3 controller, but since the load balancing mechanism is autonomous and self-regulating, we assume that the role of the RTC/C3 controller is to compensate for possible slow adaption of the distributed mechanism to sudden changes in the load distribution, and to enforce policies on the trade-off between balancing performance and cost. The RTC/C3 will thus be in control of 1) *which probabilistic metric to balance*, i.e. a running mean or a load level risk estimate, and 2) *how to map the computed "imbalance" measure to a CRE range*. The produced metrics, especially the overload risks should be useful also for other purposes, e.g. troubleshooting, radio resource allocation and deployment planning.

The method is applicable for intra-RAT load balancing for other RATs, such as Wi-Fi. Load balancing at inter-RAT level using risk as abstraction in a network graph requires inter-control communication protocols and RAT-specific mappings on to low-level parameters with respect to e.g. bias values. Implementation of an inter-RAT mechanism is out of scope for our work in WP3, but we explore the technical requirements for using the load balancing mechanism in a Wi-Fi intra-RAT setting and for showing how inter-RAT (e.g. Wi-Fi & LTE) load balancing conceptually can work in [26], and briefly summarise these results in section 4.1.3.7 below.

4.1.3.4 Logical representation

Figure 4-6 shows a conceptual outline of the network elements and information and control flow in a single RAT (e.g. LTE) setting. The diagram is RAT-agnostic and applicable to both LTE and Wi-Fi intra RAT load balancing. The NIF managers shown in the network elements are relay points for monitoring information, in this case, primarily the metrics produced by the serving nodes, and the information derived from that in the RTC/C3 controller. Downward pointing arrows indicate control parameter adjustments derived by the RTC/C3 controller(s) and distributed to the local node controllers.

4.1.3.5 Abstracted parameters

Each participating node *i* (eNodeB, or a specific Wi-Fi AP), or its RTC must be able to produce, as input to the mechanism, a normalised ([0,1] range) momentary load measurement m_i at time scale ts, e.g. 100ms. For WiFi, that could be $channel\ busy\ time/ts$. For LTE, it would be $consumed\ resource\ blocks/available\ resource\ blocks$ summed over the time period ts. Since the load balancing mechanism is formulated as a distributed algorithm, we have a controller for

each node operating at time scale ts. Balancing is achieved by manipulating bias values between every pair of potential serving nodes. A bias contribution for each node is computed by a function f from the difference between an estimate of a probabilistic metric $\hat{m}_i$ over the momentary measurements m_i and a target metric t_i produced for each node by a distributed computation within its neighbourhood. In addition, aggregate versions of the probabilistic metric, $\hat{m}_i$, the target metric t_i and the bias value $f(p_i, t_i - \hat{m}_i)$ are forwarded via the network information function (NIF) manager (see [3]) of the node controllers to a RTC/C3 controller responsible for several nearby nodes. The parameter p_i is an API for the RTC/C3 controller to influence the bias computation based on its more global view of the load situation in its area of responsibility.

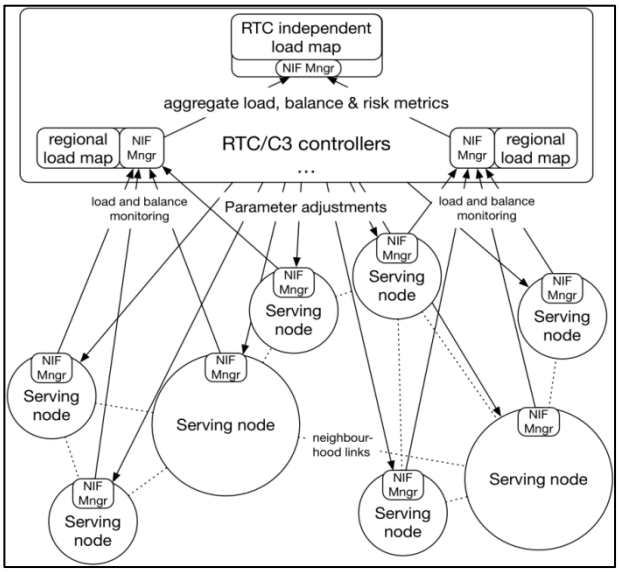


Figure 4-6 Block diagram of target system model

In its original formulation [25] the distributed algorithm used for $\hat{m}_i$ a running average over the measurements m_i, and an f that computed t_i as the average of $\hat{m}_i$ and the target values $t_i \in N_i$, i.e. in the local neighbourhood of nodes near node i. Within COHERENT, this has been generalised in two main directions:

1. Development of an alternative or complementary load metric which expresses the risk of overload or congestion at the node.
2. The introduction of the control parameter p_i to influence how the metric and target load are mapped to a bias value.
3. First steps towards RAT-agnostic load balancing and integration into a logically centralised framework.

We will focus here on the first two generalisations, and (for now) only give brief indications of how we envision further progress of the third one.

Table 4-4 Exposed parameters

Parameter	Description	Direction	Deployment
Serving node average load	Measured over a fixed sized time window, and normalised to [0,1], where 1 represents the capacity of the node	NA	RTC/R3

Serving node parametric load distribution	Estimated over a fixed number of momentary measurements	NA	RTC/C3
Risk of exceeding a given load level	Computed from CDF of above	NA	RTC/C3
Handover frequency	Running average. Useful by C3 for judging cost of achieved balance	NA	RTC

Table 4-5 Controlled Parameters

Parameter	**Description**	**Direction**	**Deployment**
Time scales	Frequency of momentary load measurements, and time window over which we estimate load distribution	NA	RTC
Overload quantile	What load level to use as threshold for risk estimates	NA	RTC
Range of biases to use	For LTE this the dB range of the CRE parameters	NA	RTC
Slope of bias mapping function	How eagerly to map actual to target metric differences to biases in the above range. Determined by C3 controller based on balancing cost metrics.	NA	RTC

4.1.3.6 Algorithms in C3

The suggested abstraction can be used in a network graph to represent the network load in terms of the estimated overload risk as a value [0,1]. The load balancing approach allows high-level requirements to be implemented and executed at different levels of the system. In the following we show two variants, in the first option Figure 4-7, high-level control parameters from the C3 layer are conveyed to local RTC controllers of the infrastructure where the load balancing algorithm will execute in a self-organized manner based on the suggested distributed target computation (DTC) algorithm, as described in [4]. In this case, the C3 is offloaded to the infrastructure, such that load is distributed based on in-network communication between the serving nodes relative to high-level actuator requirements. The risk of overload is exposed in terms of a network graph to higher layers of the architecture and other functions related to management and control.

In the second option, the optimization and execution of the load balancing approach takes place in the C3, where probabilistic modelling is employed to estimate the overload risk based on local monitoring of the infrastructure. The estimates are collected and aggregated into a network graph via the NB API, where nodes represent the serving entities and the risk of overload, and the links relates serving entities and clients. The load balancing is then performed by running an optimization algorithm over the network graph given the algorithmic control parameters. In the case of DTC, optimization can be done by offline simulations given the fraction of acceptable load and a cut-off value for specifying the actuator behaviours. The output from the optimization step is a list of client/server pairs which fulfils the load balancing requirements, executed by relevant management and controller functions through the SB APIs.

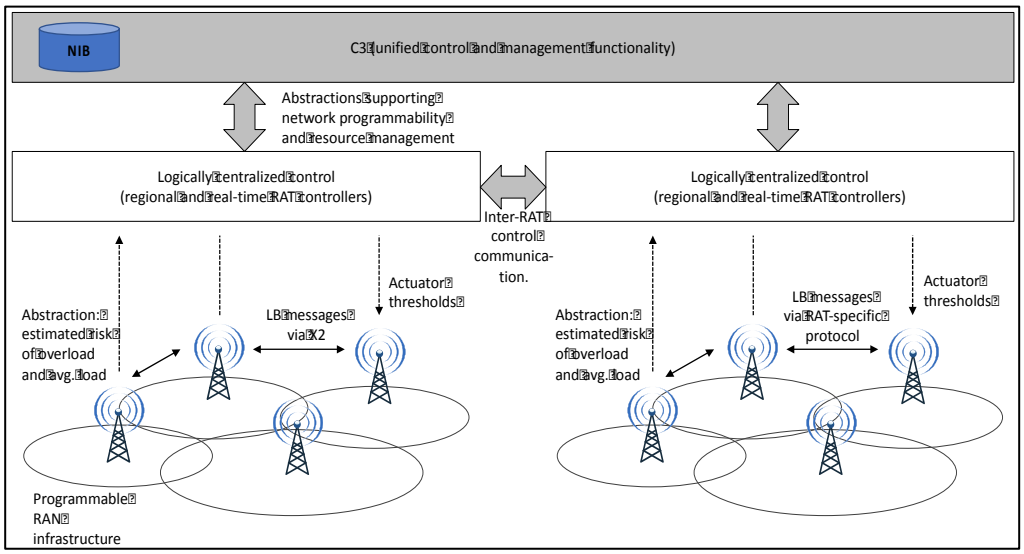


Figure 4-7 Distributed load balancing algorithm controlled by high-level requirements at C3.

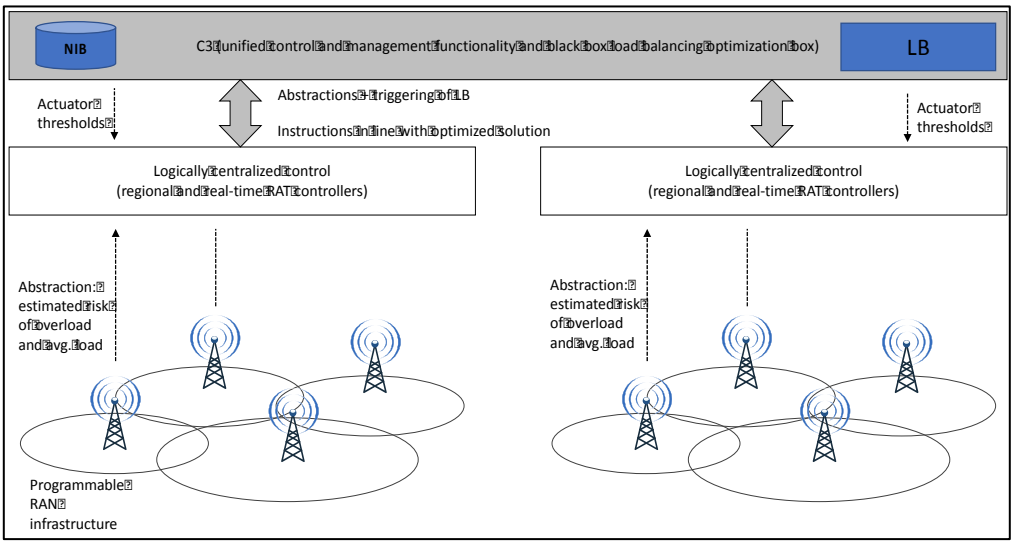


Figure 4-8 Example of centralized load balancing based on the DTC algorithm.

4.1.3.7 Verification

Extensive verification of the distributed load balancing mechanism has been performed in a stand-alone simulation framework. No implementation has yet been completed for Wi-Fi, but analysis of the requirements of the mechanism and studies of proposed schemes for handover in Wi-Fi gives strong indication that it can be implemented in a very similar way. Deployment in LTE network involves implementing the distributed target computation in base stations, initially through a proprietary extension of a node-to-node communication protocol (probably X2). The metrics computation and actuation mechanism are most easily implemented within the node controllers. Alternatively, the entire mechanism can also be implemented within a controller in charge of several nodes, but this will have a scalability impact.

Deployment in WiFi requires a similar effort on the server side but also additions or modifications of the hand-off decision mechanism within the client terminals. The following points outline the most straightforward implementation of DTC in a WiFi setting, with indications of required extensions of existing mechanisms.

- The 802.11k mechanism for generating neighbourhood reports based on RSSI probe statistics is extended to generate local neighbourhoods for the APs based on transition statistics, as in the LTE case. Alternatively, neighbourhoods may be statically configured.
- The facility for sharing link state information between APs in 802.11k is extended with the metric and target value exchanges. Alternatively, metric values, or the statistics needed to compute them, are collected at the controller and used by a centralized version of the algorithm. The metrics can be the same type of probabilistic load metrics as in the LTE case, but probably based on busy time per Wi-Fi channel, rather than cell resource blocks usage statistics.
- Computed bias values are distributed to clients via an extension of the 802.11k neighbourhood report.

The clients actuate the rebalancing operation by sending an 802.11v transition request whenever its RSSI+bias becomes less than that of a neighboring AP, according the received reports. Alternatively, the transition request can be sent from the controller or the AP, preceded by a neighborhood report indicating the desired target AP(s).

4.1.3.8 Results

Evaluation of four different variants of the mechanism is reported in [26]. These variants cover a range of configurations that would typically be under the control of an RTC/C3 controller in a fully integrated system. We present here summarized results and conclusions.

The experiments were done in an LTE HetNet setting using randomly placed nodes of different sizes, and using manipulation of the cell range expansion (CRE) parameter as the rebalancing actuation mechanism. Two types of metrics were examined: 1) a running mean over normalised load measurements m_i, and 2) an estimate of the risk of exceeding 95% of the node radio resource bandwidth. The difference between the measured metric value and the target metric is mapped into a CRE parameter settings in two distinct ways: 1) so that we exploit the range of allowed CRE values fully, and 2) using a sigmoid mapping of the range of differences into a specified range of CRE-values.

Figure 4-9 shows the distributions of changes from a base-line case with no balancing for each of seven evaluation metrics: 1) The mean load over the cells in the simulation, 2) the corresponding maximum load over the cells, 3) the load running mean fairness (Jain index), 4) the fairness over the nodes for the 95th running load percentile for the individual node, 5) the proportion of the requested bit served, and 6-7) the mean and max over the nodes for the 95th running load percentile. Each box plot represents the distribution of the respective evaluation metrics for four method variants: Load S: *Load mean* balancing metric with *full range scaling*, Load H: *Load mean* balancing metric with *sigmoid scaling*, Risk S: *Overload risk* balancing metric with *full range* scaling, and Risk H: *Overload risk* metric with *sigmoid scaling.*

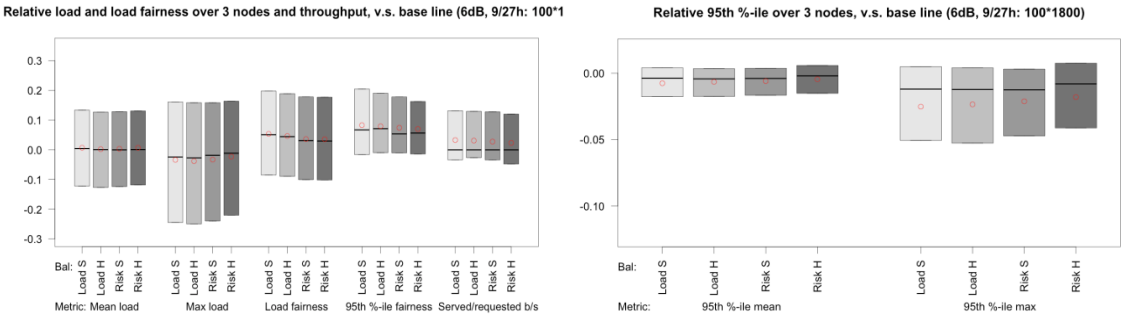


Figure 4-9 Seven evaluation metrics for four variants of the balancing mechanism for 100 scenarios with one macro node and two smaller nodes, randomly placed within the simulation area.

We can see that the mean load goes up between ~1-2% for all variants of the methods, and that this is matched by a corresponding increase in the total number of bits served. We also see that max load goes down, meaning that the most heavily loaded node over all simulation instants decreases by 1-3%. The fairness goes up significantly for both load running mean, and especially the 95:th load percentile. The load balancing method based on the risk metric efficiently reduces the occurrence of high loads, but limits the balancing effort, and cost, for cases where the over-all load is low. It is therefore preferable to the load running mean metric proposed in [25], for the autonomous balancing decisions.

All versions of the balancing method have significant and positive effect on the metrics examined, but there are also significant differences between the methods. Balancing the risk metric, as expected, achieves more balance between the very high loads, and a bit less balance overall, but this is also reflected in lower cost in terms of handovers. Highly significant reductions in handover cost can be achieved by using the sigmoid mapping, especially if we use low scaling factors and allow rare cases of extreme imbalance to result in occasional CRE offsets larger than 6dB.

The formulation of the mechanism as a distributed algorithm enables a range of implementation possibilities from completely distributed, with only sparse monitoring information and control parameter adjustments passed between controllers and nodes, to a physically centralised solution where high resolution measurement, are regularly collected and centrally managed. Regardless of the deployment strategy, load data can be aggregated to a centralised monitoring function which is a highly useful feature inherent in the design of the method, offering great implementation and operational flexibility. Furthermore, the self-adjusting balancing method offers the right type of high-level policy parametrisation to fit into a centralised management scheme using probabilistic metrics and thresholds, and essential to the use case scenarios (i.e. 1.RRS in D3.1 [4]).

4.2 Multi-connectivity and Multicasting

The functionality offered by multi-connectivity is an important tool to provide consistent user experience in heterogeneous networks consisting of multiple cells, multiple communication bands, etc. To realize multi-connectivity, resource allocation problems for both the radio access network and the backhaul network need to be addressed. Multi-connectivity function highly relies on coordination among multiple network elements (e.g. multiple BSs or multiple frequency bands), and in order to realize such a network control function, the network graph approach with the application of a centralized algorithm is a strong technology candidate. In this section, we consider two kinds of multi-connectivity functions, one is multi-cell connectivity for end-user in a single RAT LTE network, and the other is multi-RAT multi-connectivity regarding both mmWave and traditional sub-6 GHz bands which would be an interesting problem in the upcoming 5G heterogeneous mobile networks.

4.2.1 Multi connectivity in LTE

4.2.1.1 Motivation and goals

5G communication system is expected to meet the insatiable and heterogeneous demands for service and it aims to provide ultra-high bandwidth, ubiquitous super-fast connectivity and comprehensive QoS to fulfill both the end-users and service requirements. However, current mobile network deployments are usually only providing single cell connectivity, which sets certain bounds on user performance and restricts the service quality. With single-cell connectivity, the end user is only allowed to transmit and receive data with only one BS as the serving cell. To overcome this limitation, a considered approach is to utilize simultaneous multiple-cell connections for each end-user, referred as *multi-connectivity*. In general, the multiple connections can be applied among multiple RATs or within a single RAT, which is viewed as to establish multiple connections to different BSs. We focus on the LTE RAN and describe a network graph approach for the multi-connectivity problem that can be applied in the particular use case U3.FS "Flexible resource sharing for broadband PMR networks" and U6.MR "Delivery of services in public or private transportation in urban area" of D2.1 [2]. Despite the appealing of multi-connectivity, a significant challenge is to achieve efficient resource utilization among both air-interface and backhaul network. The resource utilization is crucial to enhance the user performance to secure the QoS requirement of each traffic flow across multiple connections. In general, we consider the resource allocation for two types of application traffic: (i) user-to-user such as content/video dissemination, peer-to-peer gaming and public safety, and (ii) user-to-network and network-to-user such as social networking and video-on-demand.

4.2.1.2 Requirements

Firstly, we use a bottom-up approach that abstracts the atomic parameters to formulate the required complex parameters. For instance, the underlying RSRP and noise power variance are utilized to formulate the complex parameters as SINR and achievable rate. Then, all atomic/complex parameters are used to form a network graph as the input for the high-layer RAT-agnostic algorithm executed centrally following the COHERENT hierarchical architecture. Afterward, a top-down approach based on the RAT-agnostic output will make the underlying RAT-specific decision, e.g., resource allocation, user association, backhaul capacity provisioning, etc. For instance, the RAT-agnostic algorithm allocates the user-specific data rate from connected BS to a specific user that will be translated into the PRB number and MCS index of LTE RAT.

4.2.1.3 Generation and usage of Network graphs

We use the standard notation and a network graph $\mathrm{G}(V, E)$ is constructed as following:

- ***Nodes***: The nodes represent the network entities of the heterogeneous network. All the nodes belong to the set of the graph vertices V can be either BS or user equipment (UE).
- ***Edges***: The edges (v, v'), $\forall\, v, v' \in V$ are links between the network nodes representing their relationships. The edges are considered to be directed and the ordered pair of vertices (v, v') implies the direction of the edge. All the edges belong to the set of the graph edges E.
- ***Properties***: Each directed edge contains a property depending on the use case and on which network graph is described. A property $f_{v,v'}$ is a set of K network parameters $p_{v,v'}^{(j)}, \forall\, j \in \{1, \cdots, K\}$ for the edge, defined as: $f_{v,v'} = \left\{p_{v,v'}^{(1)}, \cdots, p_{v,v'}^{(K)}\right\}, \forall\, v, v' \in V$. If the set is empty, i.e., $f_{v,v'} = \emptyset$, then there is no connectivity between the ordered pair of vertices (v, v') and thus no relevant parameters exist. Figure 4-10 shows an example of the network graph with 3 BSs and 3 UEs.

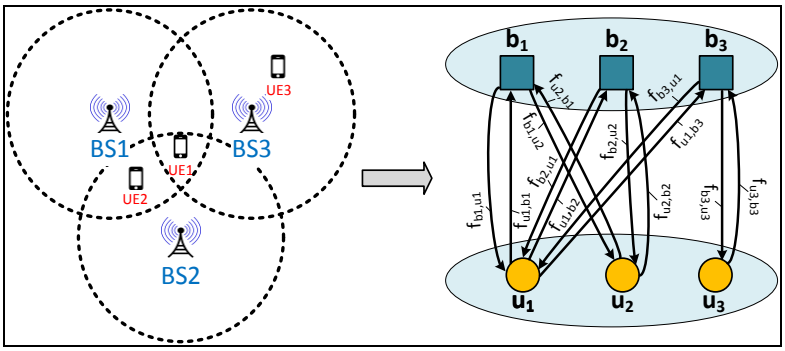


Figure 4-10 From Original network to Network graph, that represents the physical connectivity formed at the abstraction layer by extracting the radio access layer parameters

The network graph is generated based on the atomic/complex parameters updated by NIF via the graph tools using construction and update command. Further, the generated network graph is given as an input to the control applications as well as the QoS requirement of each traffic flow. Then, the result of the optimization problem is to allocate the resources for each traffic flow. This step of the algorithm is RAT-agnostic and can be run in the C3 (see Figure 4-11). Afterwards, the control functions at RTC translate the technology-agnostic allocated data rate to the RAT-dependent parameters for the underlying network. This approach can naturally apply both southbound and northbound APIs that is already within the SDN controller and coordination

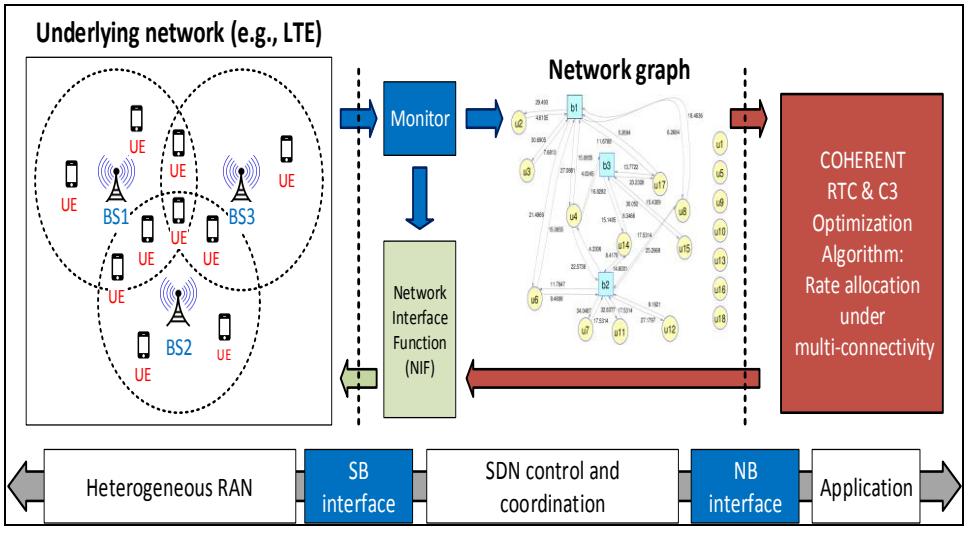


Figure 4-11 from heterogeneous RAN to the higher-layer application

4.2.1.4 Abstracted parameters

Table 4-6 Exposed parameters

Parameter	Description	Direction	Deployment
$RSRP_b$, $RSRP_u$	Reference signal received power of BS/UE	UL/DL	RTC
$SINR_b$, $SINR_u$	Signal to interference plus noise ratio of BS/UE	UL/DL	C3
B_b	Total PRB number of BS	UL/DL	RTC
B_u	Maximum PRB number of UE	UL/DL	RTC

W	Bandwidth per PRB in Hz	UL/DL	RTC
N_0	Thermal noise density	UL/DL	RTC
$\widehat{R}$	Traffic flow requested rate	UL/DL	C3
P_u	Transmission power per PRB of UE	UL	RTC
$P_{u,max}$	Maximum transmission power of UE	UL	RTC
C_h	Provisioned backhaul capacity	DL/UL	RTC

Table 4-7 Controlled parameters

Parameter	**Description**	**Direction**	**Deployment**
R^D, R^U	Achievable physical data rate of DL/UL	UL/DL	C3
x^D, x^U	Allocated PRBs to each traffic flow of DL/UL	UL/DL	RTC
MCS^D, MCS^U	Allocated MCS index of DL/UL	UL/DL	RTC

4.2.1.5 Algorithms in C3

We allocate resource to traffic flows optimally based on the applied utility function. Here we consider two utility functions: (a) *Proportional Fairness* (PF) that exploits the logarithmic utility function to maximizes the network aggregated throughput to achieve "proportional fair", and (b) *Utility Proportional Fairness* (UPF) that takes the logarithm of the sigmoid function based on QoS (in terms of the requested data rate) into account as in the following equation where y and β are the allocated data rate and the requested rate. If the allocated data rate y is less than the requested rate β, then the sigmoid function provides monotonic increment on its slope whereas it monotonic decreases in the slope when the requested rate is achieved (i.e. $y > \beta$). Moreover, γ impacts the shape of sigmoid function to be more like step or linear function reflecting the demand degree.

$$S(y, \gamma, \beta) = \frac{1}{1+e^{-\gamma(y-\beta)}} \qquad (4\text{-}3)$$

Moreover, several constraints are needed: (i) The number of allocated resource of each BS for all users shall not exceed the total number of its resources, (ii) The total number of allocated resources to each user through all connected BSs cannot exceed the capability of such user (in terms of the maximum number of allocated resources), (iii) The provisioned capacity of each backhaul link cannot be exceeded by the accumulated date rate, and (iv) Transmission power of each user to all connected BSs cannot exceed its maximum power. Moreover, the signal to interference plus noise ratio (SINR) threshold is used to represent the minimum requirement on the received signal quality to establish the connection, i.e., no connection between two nodes if the SINR threshold is not fulfilled in neither uplink nor downlink direction. The overall problem is transformed as a convex optimization problem and be solved with convex optimization problem solver to provide a unique tractable optimal solution. Such solver is executed iteratively and collocated at the C3 entity.

4.2.1.6 Results

Local-routed user-to-user traffic

We consider the user-to-user traffics that are local-routed, i.e., both users are connected to at least one common BS. Firstly, we compare the single and multi-connectivity cases both using PF utility function as shown in in different loading scenarios.

Table 4-8 in terms of average number of connected BS, number of connected user pairs, and aggregated user rate for three scenarios, namely under-loaded, uneven-loaded and over-loaded networks. The number of connected BS per UE is increased with multi-connectivity especially for under-loaded scenario since the induced inter-cell interference is minor. We observe a higher number of connected user pair in multi-connectivity as each UE is able to transmit and receive

toward more UEs across different BSs. This advantage becomes significant in over-loaded scenario allowing traffic diversity. When comparing the aggregated user rate, the performance gain is higher in under-loaded scenario due to scheduling UEs across all available BSs that is one of the expected merit of multi-connectivity. To sum up, the multi-connectivity not only has advantage in user perspective (i.e., more UEs can be reached through multiple BSs) but also in network perspective (i.e., larger aggregated user rate) in different loading scenarios.

Table 4-8 Comparison of Single/Multiple-connectivity

UE number in BS (b1/b2/b3)	Performance metric	Single-connectivity	Multi-connectivity
Under-loaded case (2/2/2)	Connected BS	1	2.07
	Connected UE pairs	6	17.52
	Aggregated user rate	0.99 Mbps	20.04 Mbps
Uneven-loaded case (2/2/2)	Connected BS	1	1.34
	Connected UE pairs	34	49.95
	Aggregated user rate	11.68 Mbps	46.73 Mbps
Over-loaded case (2/2/2)	Connected BS	1	1.45
	Connected UE pairs	90	162.22
	Aggregated user rate	55.04 Mbps	57.01 Mbps

Second, we present the results of both UPF and PF in a scenario with 4 UEs that are initially distributed to each BS in terms of two metrics: (i) QoS satisfaction ratio and (ii) unsatisfied normalized error. Table 4-9 presents the results with five different requested data rates for all user-to-user traffic flows. In terms of the satisfaction ratio, the UPF is better than the PF one in every case. Further, UPF reduces the unsatisfied normalized error by allocating resources as close as possible to the requested rate.

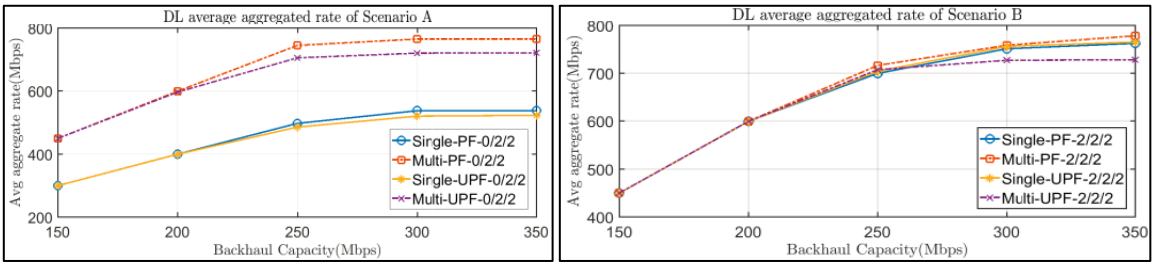
Table 4-9 QoS metric comparison of PF and UPF

Metric	Requested rate	PF function	UPF function
Satisfaction Ratio	0.1 Mbps	68.72%	91.39%
	0.5 Mbps	42.07%	58.03%
	1 Mbps	25.14%	35.33%
	5 Mbps	6.42%	23.10%
Unsatisfied normalized error	0.1 Mbps	0.2122	0.0625
	0.5 Mbps	0.4131	0.2317
	1 Mbps	0.5483	0.3906
	5 Mbps	0.7949	0.6607

Backhaul-routed user-to-network/network-to-user traffic

Then, we survey the impact of backhaul limitation. Firstly, the comparison is over the fixed and finite backhaul capacity of a start-topology backhaul network, i.e., each BS is connected to the gateway of core network via a dedicated and directed connection. In Figure 4-12 (a), the average aggregated rate is shown under single and multi-connectivity of different backhaul capacities for both UPF and PF where the a/b/c notation refers the number of UE initially distributed within three BSs. The multi-connectivity outperforms the single one. In Figure 4-12(b), there is no significant difference between single and multi-connectivity. However, the multi-connectivity can better reflect the trade-off between different utility functions as PF have a higher aggregate rate than single one and UPF has a higher satisfaction shown in Table 4-10. In Table 4-10, the satisfaction is increased in two manners: (a) use UPF function and (b) apply multi-connectivity.

Furthermore, the time-varying backhaul scenario is surveyed and it matches the wireless backhaul condition deployment. In Figure 4-13, a time-varying uniform distributed backhaul capacity between 150 Mbps and 350 Mbps is applied. We observe that multi-connectivity provides a higher aggregated rate (Figure 4-13(a)) and better user satisfaction ratio (Figure 4-13(b)) when comparing with the single-connectivity case. It strengthens our claims that multi-connectivity can provide enhancements in both network and user perspectives. Lastly, Table 4-11 summarizes the resource utilization ratio of (a) the percentage of allocated PRBs on air- interface and (b) the percentage of allocated data rate on backhaul link. The multi-connectivity better utilizes all available resources in both aspects.



(a)Scenario A with 0/2/2 UE distribution (b) Scenario B with 2/2/2 UE distribution

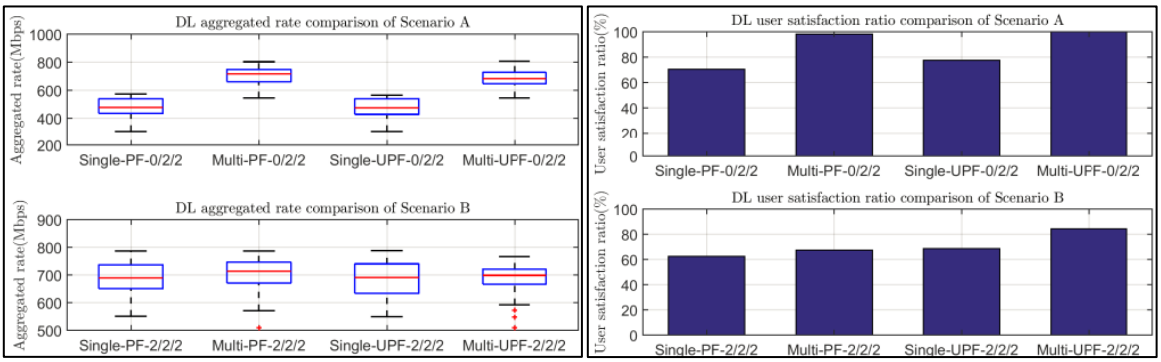
Figure 4-12 Fixed backhaul capacity impact on average aggregate rate

Table 4-10 User satisfaction ratio (%) of different BH capacity

Backhaul Capacity (Mbps)	Single-connectivity PF function		Multi-connectivity PF function		Single-connectivity UPF function		Multi-connectivity UPF function	
	A (0/2/2)	B (2/2/2)	A (0/2/2)	B (2/2/2)	A (0/2/2)	B (2/2/2)	A (0/2/2)	B (2/2/2)
150	~20	20	63	~20	~20	~20	81	~20
200	49	50	95	50	51	54	100	51
250	72	65	95	68	83	80	100	91
300	81	72	95	76	90	84	100	93
350	81	72	95	76	90	84	100	93

Table 4-11 Resource utilization ratio (%) of time-varying BH capacity

Available Resource	Single-connectivity PF function		Multi-connectivity PF function		Single-connectivity UPF function		Multi-connectivity UPF function	
	A (0/2/2)	B (2/2/2)	A (0/2/2)	B (2/2/2)	A (0/2/2)	B (2/2/2)	A (0/2/2)	B (2/2/2)
Air-interface	57	90	94	98	57	90	93	98
Backhaul	63	92	95	93	63	92	93	91



(a)Scenario A with 0/z/z UE distribution (b) Scenario B with z/z/z UE distribution

Figure 4-13 Time-varying backhaul capacity of both Scenarios

4.2.1.7 Conclusion

Multi-connectivity brings benefits both from a network and user perspective. Moreover, the UPF function satisfies QoS in terms of requested rate and increases the satisfaction ratio when there is available radio resource. The COHERENT hierarchical architecture is the enabler for the centralized RAT-agnostic rate allocation and distributed RAT-dependent resource allocation.

4.2.2 Multi-connectivity in mmWave networks and LTE

4.2.2.1 Motivation and goals

Future 5G mobile networks are envisioned to exploit both conventional sub-6 GHz frequencies (below 6 GHz) and the extreme high-frequency spectrum, such as millimeter-wave (mmWave) bands to provide ever-increasing throughput for users in a consistent way. With current cellular technologies, the service provided to users at cell edge area is relatively poor, even when applying complicated joint-processing collaborative multipoint transmission technologies. Further densification of wireless networks using mmWave bands, combined with massive multiple-input multiple-output (MIMO) and beamforming techniques, provides a framework to achieve throughput in the range of Gbps. However, worse radio propagation at mmWave bands leads to a requirement to densify the network just to keep the relative cell edge performance constant. To achieve consistent user experience [42] would require even further network densification. To provide continuous and reliable services in 5G mmWave networks, multi-connectivity [43], [44] is an essential function for mmWave networks. First, mmWave communication is susceptible to the blockage effects, mobile users are expected to experience high fluctuation in throughput performance if they are connected to one single mmWave BS [45]. Second, compared to traditional low frequency bands, the coverage area of mmWave BS is limited and discontinuous due to the lack of diffraction and the high pathloss of high frequencies due to small antenna aperture [45], [46], [47]. MmWave band alone cannot provide continuous services except that the mmWave BSs are deployed in an ultra-dense way according to the environments by using advanced planning methods based on real measurement campaigns or ray tracing simulations. Using multi-RAT multi-connectivity technologies with both sub-6GHz and mmWave bands will ease the deployment requirements for mmWave and reduce the mmWave deployment cost.

In this section, we study the multi-connectivity problem in a 5G multi-RAT network with both mmWave BSs and traditional sub-6 GHz BSs (e.g. a LTE BS). This application is directly linked to the RAT sharing use case (UC1.SR) described in D2.1 [2]. Based on the abstracted parameters, a multi-connectivity network graph is used to characterize the multi-RAT connectivity relationship between multi-RAT UEs, mmWave BSs, sub-6 GHz BSs or the integrated sub-6GHz & mmWave BSs. Using the multi-connectivity network graph, a centralized algorithm for cell association is

performed by C3 and a multi-RAT resource allocation algorithm is performed by sub-6 GHz/mmWave BS RTC, to increase a KPI called User Experience Consistency (UEC), which is defined as the ratio of cell-edge user throughput and the network mean throughput.

4.2.2.2 Requirements

The introduction of multi-connectivity function in 5G mmWave network, and the heterogeneity of frequency resources (i.e. sub-6GHz and mmWave resources) will complicate the network management for a multi-user multi-cell network system. Both sub-6GHz BSs and mmWave BSs need to be shared by multiple users. Sub-6GHz connections (i.e., radio resources) should be allocated to users who cannot find a proper mmWave BS. In order to provide globally optimal BS association strategy and perform resource allocation for the abovementioned mmWave networks with multi-connectivity function, a centralized control framework is required to coordinate the multiple BSs. UEs and BSs need to perform mmWave beam discovery and channel measurement periodically. A local RTC is responsible to collect the raw channel measurement results and estimate the capacities of related connections. These estimations will be sent to the central controller (C3) and C3 will perform multi-connectivity algorithm based on a multi-connectivity network graph derived from the estimates for all connections including mmWave connections and sub-6GHz connections by C3.

4.2.2.3 Generation and usage of Network graphs

We envision three kinds of multi-connectivity scenarios in future heterogeneous 5G wireless networks, as shown in Figure 4-14.

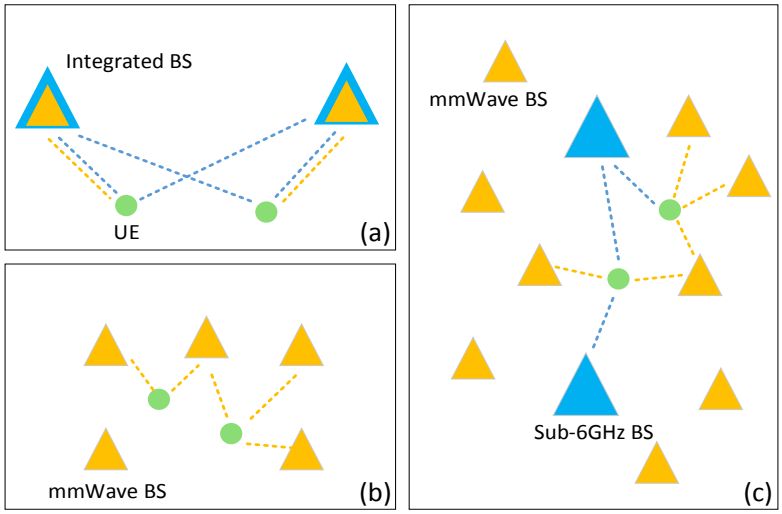


Figure 4-14 Multi-connectivity in mmWave networks and the network graphs

First, a cheap and reasonable way to introduce mmWave technology into cellular network is to tightly integrate a mmWave network with an existing sub-6GHz network, which results in an integrated network [48]. The sub-6GHz and mmWave BSs are co-located as shown in Figure 4-14(a). Second, for stand-alone mmWave ultra-dense networks (UDN) in Figure 4-14(b), multi-connectivity function can be run on multiple mmWave BSs to provide multiple mmWave connections to a user. In this case, the same user is connected to multiple mmWave BSs by using directional beamforming transmission from each mmWave BS. A local real-time controller (RTC) is responsible to control the local mmWave BS cluster, and a central controller (C3) is responsible to choose a set of mmWave BSs for a user based on the measurements statistics. The redundancy of connections can increase the reliability as UE can switch to another mmWave BS using a new beam if the current used one is blocked on its direction. Different from traditional cellular

handover, the switching is performed only on PHY layer and does not require high layer signaling overhead. Third, future mmWave networks can be deployed under the umbrella of a traditional low-frequency (sub-6 GHz) macro or micro cellular network and is controlled by this underlay network. In such a sub-6 GHz + mmWave two-layer network as shown in Figure 4-14(c), a user is always connected to a sub-6GHz BS, this connection provides control-plane information and reliable data transmission for users when no mmWave connection is available. At the same time, a user can have multiple connections on mmWave layer just as the case in stand-alone mmWave UDN.

The sub-6GHz band is narrow but has reliable and continuous coverage. The mmWave band is wide but has unreliable, discontinuous coverage. The sub-6GHz and mmWave carriers thus complement each other in a multi-connectivity phantom cell concept. There are two kinds of frequency resources in a sub-6GHz + mmWave network (networks in Figure 4-14 (a) and (c)), the precious and reliable low-frequency resources which can provide continuous performance for users and the abundant and unreliable mmWave resources which can only provide high performance for closed users or line-of-sight users. For network (a), user equipment (UEs) are connecting to the integrated base-stations based on their channel qualities to the BSs on both sub-6GHz and mmWave bands; both sub-6Ghz and mmWave resources are allocated to the same users considering its instantaneous channel measurements on these two separated bands. It should be noticed that one user may be associated to two BSs at the same time, one with the best sub-6Ghz channel and one with the best mmWave channel.

Network graph is used to characterize the multi-connectivity relationship between users, sub-6GHz BSs and mmWave BSs. The multi-connectivity NIF will estimate the short-term average link capacities for all connections based on the channel measurements and construct the weights of the graph edges using such estimated statistics. A user (or BS) would have edges connected to multiple BSs (or users). Based on the multi-connectivity network graph, the network scheduler can perform sub-6GHz & mmWave resources allocation for multiple users considering some utilities such as sum capacity or fairness. The network scheduler is hierarchical, and its function will be run in both C3 and RTC. C3 is responsible to find a subgraph in the multi-connectivity network graph that is utilized to determine which BSs are actually connected to a user. While the C3 is to determine the connection relationship, the RTC is a local agent to provide fast resource allocation for the users among multiple neighboring BSs. The C3 can also perform load balancing function based on the multi-connectivity network graph when determining the actual connection relationship, by shifting user from congested BSs to under-loaded BSs.

4.2.2.4 Abstracted parameters

The measurements required for the considered multi-connectivity applications are the traditional sub-6GHz channel measurements (such as CQI, PMI and RI in LTE, RSSI in 802.11) and the beam-based or pilot-based channel measurements in mmWave with large antenna array. To reduce the signaling overhead between RTC and C3, a network information function need to be carefully designed to translate these raw measurements for both sub-6GHz and mmWave into reliable and accurate link capacity estimation. This network information function could be a probabilistic method utilizing historical information. A simple NIF is the Shannon mapping from average SINR to link capacity. When joint-processing collaborative multipoint transmission is enabled, the NIF also need to estimate the virtual link capacity between the user and the multiple collaborative mmWave BSs. In this case, the vertex (node) in a multi-connectivity network graph can be a set of co-located mmWave BSs. An advanced NIF is one that can predict the link capacity based on the current and past measurement results. Using some context information, such as positions and velocities of moving users as well as tome historical measurements, the multi-connectivity NIF can give good predictions for link capacities and ensure that proper set of mmWave BS candidates are selected for a user. Table 4-12 and Table 4-13 show the abstracted parameters to be used in this application.

Table 4-12 Common exposed parameters

Parameter	Description	Type	Direction	Deployment
$\sigma_{i,j}$	Connectivity between node *i* and *j* , $\sigma_{i,j} = 1$ if $SNR_{i,j} \geq SNR_{th}$, and $\sigma_{i,j} = 0$ if $SNR_{i,j} < SNR_{th}$, where SNR_{th} is the minimum SNR required for a data transmission.	Complex	UL/DL	RTC/C3
$\xi_{i,j}^{(T)}$	Link reliability between node *i* and *j*, it is defined as the packet error rate (PER) during an evaluation time period *T*.	Complex	UL/DL	RTC/C3
$m_{i,j}^{(T)}$	Mean throughput between node *i* and *j*, during an evaluation time period *T*, using one frequency resource block.	Complex	UL/DL	RTC/C3

Table 4-13 Common controlled parameters

Parameter	Description	Direction	Deployment
$c_{i,j}$	$c_{i,j}$ = 1 if UE *j* is associated with BS *i*, and BS *i* will allocate resources to UE *j*, otherwise $c_{i,j} = 0$, and no resource will be allocated to link (*i*, *j*).	DL	C3
$N_{i,j}$	Number of time/frequency resources allocated to link between BS *i* and UE *j*.	DL/UL	RTC

4.2.2.5 Algorithms in C3

The C3 is the central coordinator which collects global network state information reported by local BSs and controls the multiple BSs by sending control commands as shown in Figure 4-15. The information collected from the BSs includes: (a) the number of connected UEs; (b) the traffic load inside the cell; (c) link capacity predictions for all connections. The control functionalities in C3 include: (a) definition of utility function for resource allocation; (b) multi-connectivity algorithm; (c) load balancing algorithm. The RTC is located at each BS, and is responsible for (a) perform sub-6GHz & mmWave resource allocation for the connected users based on the scheduling metric set by the C3; (b) Collecting information and channel measurements from UEs and BSs, and do the link reliability and mean throughput estimation, and report them to C3.

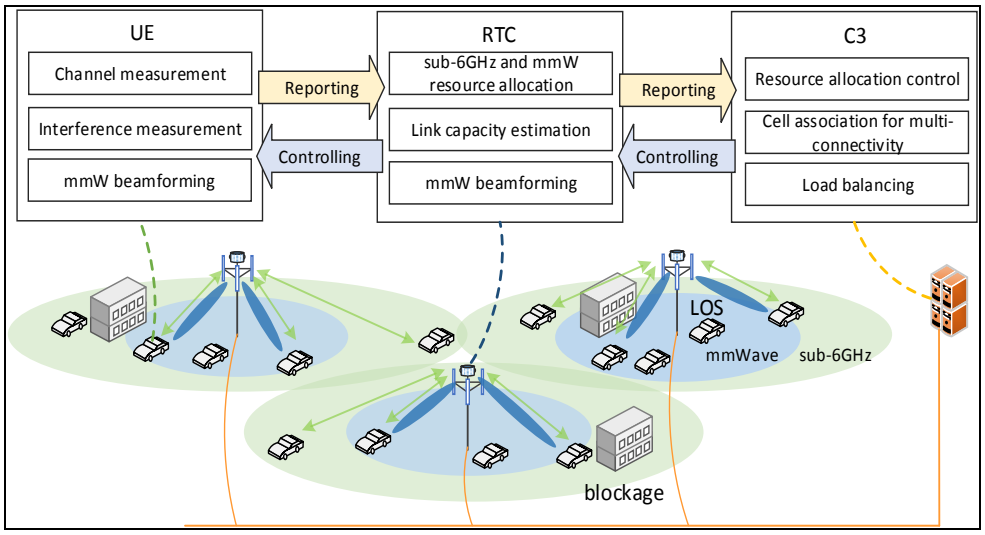


Figure 4-15 Control framework for sub6-GHz & mmWave multi-connectivity

Assuming that the sub 6-GHz has B_1 resource blocks and mmWave has B_2 resource blocks, there are totally N UEs, and M_1 sub6-GHz BSs and M_2 mmWave BSs in the network. In the integrated network, we will have $M_1 = M_2$. After channel measurements and reporting, the C3 will have a connectivity network graph represented by a connectivity matrix $[\sigma_{i,j}]_{N\times(M_1+M_2)}$, and a mean throughput matrix $[m_{i,j}]_{N\times(M_1+M_2)}$. We assume all UEs are fully-loaded with full-buffer traffic, and the object function of network control is proportional fairness for all UEs. To maximize the global objective function, we use a hierarchical scheme. First, C3 tries to find an optimal cell association matrix$[c_{i,j}]_{N\times(M_1+M_2)}$ by assuming that both mmWave and sub-6GHz resources for a local BS are equally allocated to the associated UEs with this association matrix. The C3 will send this cell association matrix to the local RTCs, each will then try to maximize a local objective function for all the UEs associated to it, under the local radio resource allocation constraint. To maximize local objective function in RTC, resource numbers for the associated UEs are assumed to be a continuous positive real number, and gradient descent algorithm is used to find the optimal solution for local resource allocation. The detail of the algorithms is presented in D5.2 [7].

4.2.2.6 Results

In this section, we study the performance gains achieved via multi-connectivity and the proposed control framework in Manhattan scenario for an integrated sub-6GHz & mmWave network. One can find the details of the simulation parameters in D5.2 [7]. To illustrate the benefit of multi-connectivity in future mmWave networks, we consider the multi-connectivity in the integrated sub-6GHz & mmWave network. A proportional fairness utility is assumed, the network scheduler allocation sub-6GHz and mmWave resources to multiple users by maximizing the fairness utility under the constraints of limited resources**.** We compare the throughput CDF performance for three settings: 1) sub-6GHz network with single-connectivity; 2) mmWave network with single connectivity and 3) integrated sub-6GHz & mmWave network with multi-connectivity. Figure 4-16 shows the simulation results for three settings, in a network with Inter Site Distance (ISD) as 283 meter. Using stand-alone mmWave connection, a lot of users are in mmWave outage. As compared, sub-6GHz & mmWave dual-connectivity can increase the performance of cell edge users significantly. The performance gain is achieved by allocating more sub-6GHz resources for cell-edge UEs since UEs close to the BS would use mmWave resources for their communications.

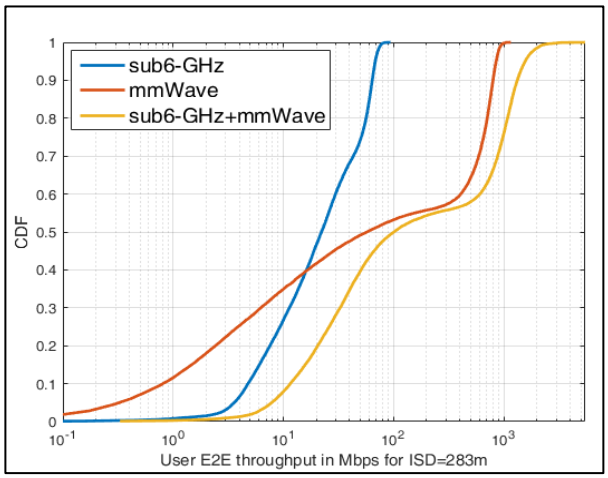


Figure 4-16 Simulation results for multi-connectivity in an integrated network

4.2.2.7 Conclusion

We have considered a hierarchical control framework to address these network management problems related to mmWave & sub-6 GHz multi-connectivity. This work was published in part in [88]. The computation overhead for network optimization is reduced by dividing the optimization problem into two levels, i.e. a centralized cell association problem and the local resource allocation problems. System level simulation in urban micro-cell scenarios illustrates that using mmWave & sub-6GHz multi-connectivity can increase the cell-edge performance significantly, which can satisfy the 5G requirement of consistent user experience.

4.2.3 Multicasting in WiFi

4.2.3.1 Motivation and goals

Wireless communications are witnessing an exponential traffic growth. Multimedia communications account for a significant fraction of that traffic. Among all the various types of multimedia communications, a particular role is played by multicast and broadcast services. In multicast/broadcast context, content is delivered to a group of users or to all the users. Due to its stochastic nature, the proper support of multicast communications in wireless and mobile networks is particularly challenging. Nevertheless, the popularity of multicast applications makes supporting such kind of use cases on wireless networks, in general, and on 802.11–based WLANs [27], in particular, of capital importance. 802.11 WLANs dynamically choose among different Modulation and Coding Schemes (MCS) for frame transmission. For example, in the case of 802.11a/g networks, devices can choose among MCS resulting in physical layer bitrates ranging from 1 to 54 Mb/s, while in the case of the 802.11n/ac networks higher throughput MCSes are also available. The rate control mechanism is however restricted to unicast transmissions. This is because 802.11 uses a two–way handshake protocol where each data transmission must be acknowledged by the receiver. However, in multicasting, acknowledgements cannot be used as they would inevitably collide at the transmitter. As a result, multicast transmissions are usually performed at the lowest MCS (to increase both the range and the reliability of the transmission) and do not use any form of transmission feedback mechanism. This has several drawbacks: (i) it severally limits throughput, (ii) it consumes more radio resources affecting also the capacity available to other (unicast) flows, and (iii) since multicast frames are not retransmitted, the reliability of the multicast streams can be adversely impacted. This work aims to improve the quality of experience of multicast communications while optimizing the utilization of radio resources. This goal is achieved by

selecting the multicast transmission rate that can deliver the expected quality of experience in terms of both throughput and frame loss ratio and by associating the multicast receptors in such a way to minimize the radio resource utilization across the entire network.

4.2.3.2 Requirements

A fundamental requirement in achieving multicast rate adaptation and mobility management is network programmability which is achieved by using dynamic reconfiguration of network control parameters. Another requirement is the collection of all parameters (real and abstracted) which are representative of the state of the network's constituent entities. In WiFi networks, these entities include the wireless stations and access points at which a large number of observable parameters can be collected. In the context of this work, we have developed and used abstractions that fulfill these afore-mentioned requirements for the network re-programmability based on its representational state to achieve multicast rate adaptation and mobility management.

4.2.3.3 Generation and usage of Network graphs

The network graph is an abstract network state representation that can be manipulated by C3 to achieve a target network state. Instantaneous measurements are collected, categorized in Rate Control Statistics and Transmission Policies, and exposed via the southbound interface to the control and coordination layer. The control layer creates the Global Network View (network graphs) and exposes it to the application doing the actual decision making. Notice how the 5G-EmPOWER has been used as reference implementation of the COHERENT C3 entity. The COHERENT SDK has been implemented on top of 5G-EmPOWER to enable network programmability. Applications run at the application layer leveraging on the global network view exposed by the controller to implement the network intelligence through the COHERENT SDK.

4.2.3.4 Abstracted parameters

Light Virtual Access Point (LVAP): The LVAP abstraction [29] provides a high–level interface for wireless clients' state management. This interface handles all the technology–dependent details such as association, authentication, handover, and resource management. A client attempting to join the network triggers the creation of a new LVAP. In Wi-Fi networks, an LVAP is a per–client virtual access point which simplifies network management and introduces seamless mobility support, i.e., the LVAP abstraction captures the complexities of the IEEE 802.11 protocol stack such as handling of client associations, authentication, and handovers. More specifically, every newly received probe request frame from a client is forwarded to the controller which may trigger the generation of a probe response frame through the creation of an LVAP at the requesting AP. The LVAP thus becomes a potential candidate AP for the client to perform an association. The controller can also decide whether the network has sufficient resources left to handle the new client and might suppress the generation of the LVAP. Each AP hosts as many LVAPs as the number of wireless clients that are currently under its control. Each LVAP has an identifier that is specific to the newly associated client. Removing a LVAP from an AP and instantiating it on another AP effectively results in a handover.

Transmission Policy: The physical layer parameters that characterize the Wi-Fi radio link, such as transmission power, MCS, and Multiple Input Multiple Output (MIMO) configuration must be adapted in real–time. Consequently, programming abstractions for rate–adaptation in Wi–Fi networks must clearly separate fast–control operations that must happen very close to the air interface, such as rate adaptation, from operations with looser latency constrains, such as mobility management. The *Transmission Policy* abstraction allows an SDN controller to reconfigure or replace a certain rate control policy if its optimal operating conditions are not met. The Transmission Policy specifies the range of parameters the AP can use for its communication with a wireless client. Table 4-14 and 4-15 summarizes the considered parameters.

Table 4-14 Exposed parameters

Parameter	Description	Direction	Deployment
RSSI	Received Signal Strength Indicator measured at the Access Point	DL	C3

Table 4-15 Controlled parameters

Parameter	Description	Direction	Deployment
Client association	Point of attachment of a wireless client	DL/UL	C3
Transmission Policy	MCS available to the rate control algorithm	DL	C3

4.2.3.5 Algorithms in C3

Multicast Rate Adaptation: The Multicast Rate Adaptation algorithm steers the multicast rate selection towards an efficient operating point. The main idea is to use the link delivery statistics collected at the AP to dynamically adapt the MCS in Legacy mode. In Legacy mode, broadcasts are used for multicasting instead of the unicast mode used in Direct Multicast Service (DMS). However, the rate adaptation algorithm, as per standard, is used only for unicast transmissions. As a result, if there are no ongoing unicast transmissions, no link delivery statistics will be computed. To circumvent this issue, we introduce the following two phased scheme. First, the controller uses the Transmission Policy abstractions to specify DMS as multicast policy for a specific multicast address. This allows the rate adaptation algorithm to kick–in and gather the link delivery statistics for all the receptors. Second, the controller uses the link delivery statistics to compute the MCS with the highest successful delivery probability for every receptor in the group. Subsequently, a worst receptor approach is used to compute the optimal transmission MCS. The whole process, sketched in Figure 4-17, is repeated periodically with a configurable ratio between DMS and Legacy periods. This allows the programmer to trade accuracy for airtime utilization. Specifically, by increasing the fraction of time of the DMS mode it is possible to improve the link delivery ratio at the price of higher channel utilization.

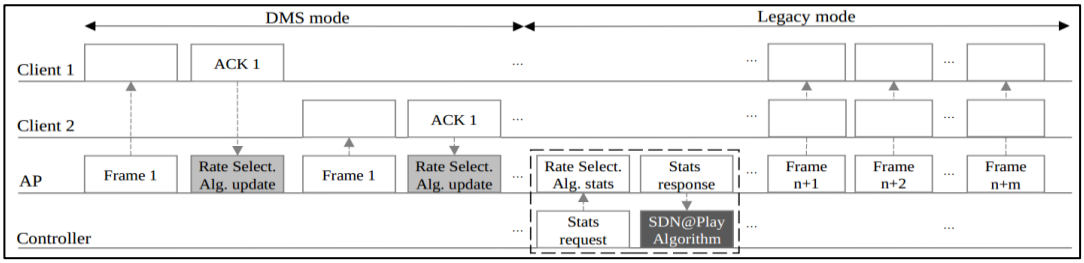


Figure 4-17 Information gathering and Multicast data rate calculation process

A. Statistics gathering: The 5G-EmPOWER platform, used as the control plane, provides a rich set of programming primitives through a Python-based Software Development Kit (SDK) [28], which can operate in either polling or trigger mode. In polling mode, the controller polls one or more APs for specific information and a thread is created with defined firing conditions. We have defined a new polling–based primitive allowing the controller to access to the rate adaptation algorithm statistic. For each supported MCS, the Exponentially Weighted Moving Average (EWMA) of the frame delivery probability and the expected throughput in the last observation window are reported. The total number of successful and failed transmissions, in the last

observation period, is also reported. This primitive is used by the control application to gather the link delivery statistics for all the wireless clients involved in multicast transmissions.

Data-path Implementation: Each AP consists of two components: one OpenvSwitch [31] managing the communication over the wired backhaul; and one Click modular router [32] implementing the 802.11 data–path. Click is a framework for writing multi–purpose packet processing engines. Communications between Click and the controller takes place over a persistent TCP connection. Rate adaptation is also implemented in Click using the Minstrel [33] algorithm. Minstrel operations follow a multi–rate retry chain model in which four rate–count pairs, r0/c0, r1/c1, r2/c2 and r3/c3, are defined. Each pair specifies the rate at which a unicast frame shall be transmitted and a fixed number of retry attempts. On successful packet transmission, the remainder of the retry chain is ignored. Otherwise the AP moves to the next pair in the chain. For each supported MCS, Minstrel tracks the link delivery ratio and the expected packet throughput given the probability of success. Statistics are recomputed every 500ms. The rates with the highest throughput, second highest throughput, and highest delivery probability are maintained. Minstrel spends part of its time in a so-called look–around mode. Specifically, 90% of the time, Minstrel configures the retry chain using the collected link delivery statistics. In the remaining 10% of the time it randomly tries other MCSes to gather statistics. Table 4-16 summarizes the criteria used by Minstrel to fill the retry chain in both normal and look–around mode. We extended the Click data–path implementation to support generalized transmission policies for unicast, multicast, and broadcast addresses. According to the new transmission policies, the rate adaptation algorithm (i.e., Minstrel) will use the first entry in the list of available MCSes if the multicast mode is set to Legacy. Conversely, if the multicast mode is DMS, the frame is duplicated for each receptor in the group and is fed back to the rate control algorithm which applies the unicast transmission policy.

Table 4-16 Minstrel Retry Chain Configuration

Rate	Look–around		Normal transmission
	Random < Best	Random > Best	
$r0$	Best rate	Random rate	Best rate
$r1$	Random rate	Best rate	Second best rate
$r2$	Best probability	Best probability	Best probability
$r3$	Base rate	Base rate	Base rate

Mobility Management: Distance and interference may affect the multicast performance and force APs to use less efficient MCSes. This, in time, may result in a higher channel utilization and higher frame loss ratio. SDN@Play Mobile aims at jointly optimizing client association and multicast rate selection while minimizing the channel occupancy. To this end, multicast receptors report to their serving AP the list of neighboring APs and the experienced channel quality. This information is gathered using the Beacon Reports of 802.11k [30]. Beacon Reports enable an AP to request a list of APs from a client whose beacons it can receive on a specified channel/s. The clients monitor beacon and probe response power levels and report the measurements in a Beacon Report. Beacon Reports are aggregated by the AP and reported to the controller which builds a downlink channel quality map. For each multicast receptor, the controller periodically checks the channel quality between the receptor and all its neighboring APs. If the signal strength of serving AP goes below −75 dBm or another AP has better channel quality by at least 20 dBm for five consecutive checks, then a handover is triggered. This reduces ping-pong handovers. After the handover, the SDN@Play algorithm re-computes the multicast transmission rate for all APs. If, as result of handover, the global channel occupancy increases, the handover is reverted. To avoid oscillations, if a handover is reverted then the target AP is not considered as candidate AP.

4.2.3.6 Results

Evaluation Methodology: The testbed configuration used for our experimental performance evaluation is depicted in Figure 4-18. It consists in four multicast receptors (MR_i), 3 APs (AP_j), one controller (C), one video server (S), and one Ethernet switch (SW). The multicast receptor MR1 is a mobile station, while the remaining three are static. The measurement campaign described in this section has been executed over one floor of an office environment. During the measurements three multicast receptors ($MR_{2,3,4}$) were kept in the same position, while one receptor (MR_1) moved along a 50m long corridor. The receptor MR_1 is initially located near AP1. Then, the receptor moves towards AP3. The corridor is divided into 10 segments. At the end of each segment the station stops for 20 seconds. This results in an average speed of 0.2 m/s. A multicast video stream is generated by the video server S and delivered to a group of receptors. The video is a five minutes HEVC encoded sequence [34] and transmitted using Ffmpeg [35]. Two different compression schemes with average bitrate of 1.2 and 6.2 Mb/s are considered. This way, it is possible to obtain detailed information regarding different bitrate affects. Three multicast strategies have been considered in this study: Legacy Multicast, SDN@Play, and SDN@Play Mobile. We considered delivery ratio and channel utilization as performance metrics. All the evaluations were taken in 11a mode, with 6Mb/s Legacy multicast rate. In the SDN@Play case, the algorithm spent 500ms in DMS and 2500ms in Legacy mode. Between each measurement the rate adaptation statistics are cleared. As no other downlink traffic is considered, the only opportunity for Minstrel algorithm to be executed is during the DMS periods

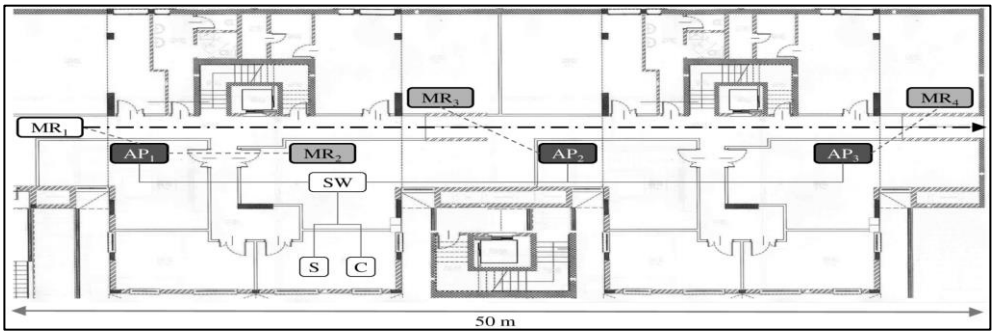


Figure 4-18 Testbed deployment layout

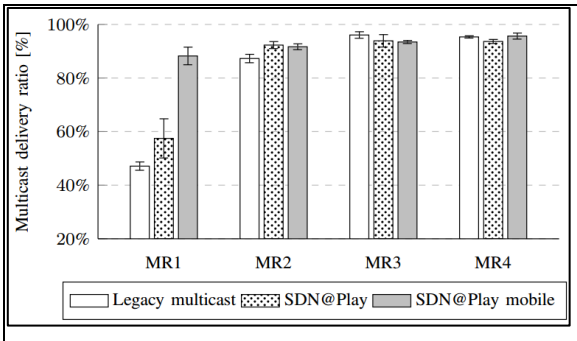


Figure 4-19 Delivery ratio for multicast video transmission at 1.2Mbps for each receptor.

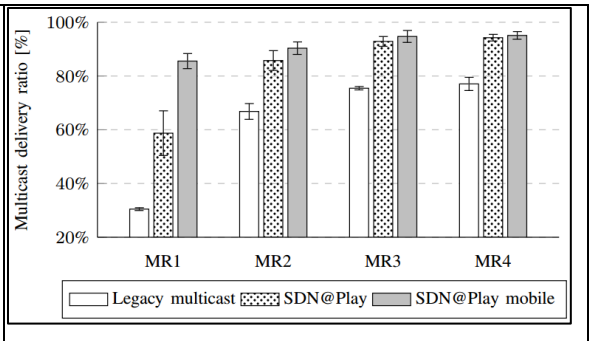


Figure 4-20 Delivery ratio for the multicast video transmission at 6.2 Mbps for each receptor

Experimental results Figure 4-19 plots the delivery ratio for multicast receivers. As shown, at 1.2 Mb/s, the performance of the static receptors (MRs 2 to 4) is not affected by the multicast scheme. Conversely, SDN@Play as well as SDN@Play Mobile provide a significant performance boost to

MR1. This is since, the Legacy Multicast scheme cannot adapt to the changing channel condition experienced by the mobile receptor. Moreover, given the absence of ACKs and retransmissions, the mobile receptor starts suffering from heavy packet losses as it moves away from AP1. The SDN@Play multicast scheme performs better than the Legacy Multicast scheme as it can adapt to the changing channel conditions. Moreover, in DMS mode, the SDN@Play scheme can still retransmit some of the lost frames. Nevertheless, SDN@Play does not support mobility and thus, the mobile station remains attached to the initial AP until the connection is lost. By contrast, SDN@Play Mobile enhances the performance because the receptor is always associated to the AP that can deliver the best performance in terms of both packet loss rate and channel utilization.

Figure 4-20 plots the delivery ratio using different multicast schemes with 6.2 Mbps video. As shown, in Legacy Multicast scheme using a video with higher bitrate results performance drop for all multicast receptors. The performance drop is particularly significant in case of mobile station which experiences a 70% frame loss ratio. Conversely, SDN@Play can improve the performance of all receptors showing a delivery rate as good as the 1.2 Mb/s video. Finally, the SDN@Play Mobile multicast scheme can improve the delivery ratio for the mobile receptor by 180% bringing it at the level of 1.2 Mb/s video. The same behavior is shown in Figure 4-21 which summarizes the average delivery ratio using different multicast schemes and bitrates. Due to the low performance achieved by the Legacy multicast and SDN@Play schemes for mobile stations, the deviation shown is much higher than SDN@Play Mobile, which indicates that all the multicast receptors receive practically the same data, regardless of their position. The fact that the Legacy Multicast scheme always uses the basic data rates results in a very high channel utilization.

As shown in Figure 4-22, the channel utilization for the 1.2 Mb/s video stream is as high as 20% while in the case of the 6.2 Mb/s stream (Figure 4-23) the utilization reaches 90% making the channel unavailable for other traffic. By using higher MCS indexes, SDN@Play can decrease the channel utilization for both the static and the mobile receptors. The channel utilization improvement is even more significant in SDN@Play Mobile. As a matter of fact, in contrast to the previous case, the SDN@Play Mobile scheme can specifically address the needs of the mobile receptor by both optimizing the channel utilization and balancing the workload across the entire network.

Figure 4-24 summarizes the network–wide channel utilization using different multicast schemes and bitrates. As shown, both SDN@Play and SDN@Play Mobile show a significantly lower global channel utilization than the Legacy Multicast scheme. Figure 4-25 plots the instantaneous channel utilization at AP1. As shown, in Legacy Multicast scheme, the channel utilization remains constant and is almost always higher than the channel utilization for the other two schemes. This is since the Legacy Multicast always uses the basic MCS. Conversely, SDN@Play can use higher MCS indexes which in time results in lower channel utilization. However, while mobile recepters, the SDN@Play scheme is forced to use lower MCS indexes in order to provide the expected QoE. Eventually, this results in using the base MCS scheme. The SDN@Play Mobile jointly optimizes MCS selection and receptor association to address this issue. As shown in figure, in SDN@Play Mobile, the channel utilization at AP1 remains constant during the entire measurement. The same considerations apply to the scenario with the 6.2 Mb/s video stream (see Figure 4-26).

However, in this case the SDN@Play scheme never reaches the channel utilization of the Legacy Multicast scheme. This is since, using the 6.2 Mb/s video stream, always results in the channel being fully utilized when the Legacy Multicast scheme is used. Finally, Figure 4-27 and Figure 4-28 report the distribution of the MCS used by each AP at 1.2 and 6.2 Mb/s.

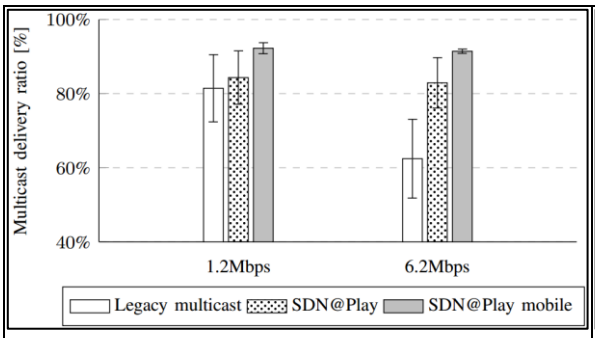


Figure 4-21 Network-wide delivery ratio using different multicast schemes at different bitrates.

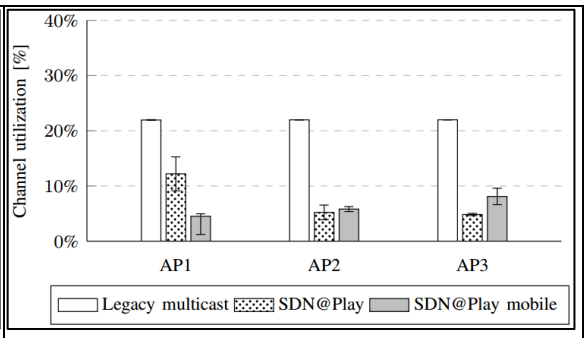


Figure 4-22 Channel utilization for the multicast video transmission at 1.2 Mbps

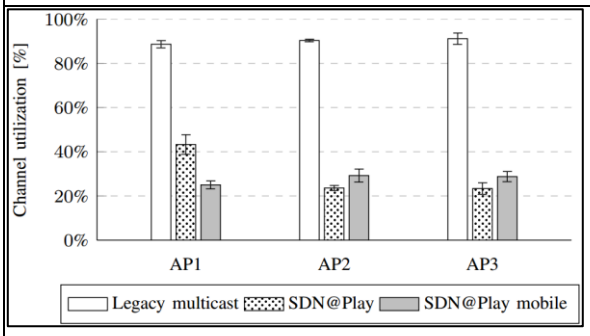


Figure 4-23 Channel utilization per AP for the multicast video transmission at 6.2 Mbps

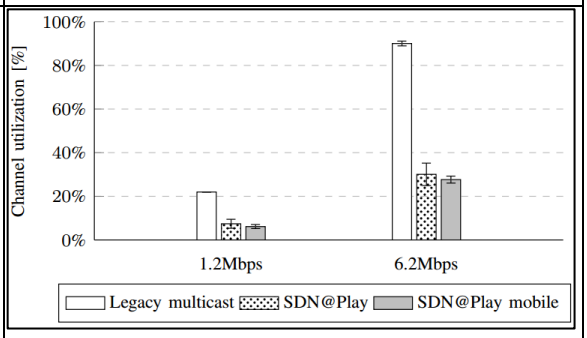


Figure 4-24 Network-wide channel utilization using different multicast schemes and bitrates

The Legacy Multicast scheme is omitted because the lowest MCS index is always used. Although especially at high bitrates SDN@Play Mobile uses high MCSes indexes in the last 2 APs for longer periods than SDN@Play, this ratio is small in comparison with the distribution obtained in AP1. In this case, it can be noticed how SDN@Play Mobile transmits 70% of the data at the highest MCS (54 Mb/s). This is since SDN@Play Mobile is able to handover the clients to AP that provides the highest network performance. On the contrary, this value is approximately the half for the SDN@Play scheme due to the distance of the mobile station from the AP it is connected to.

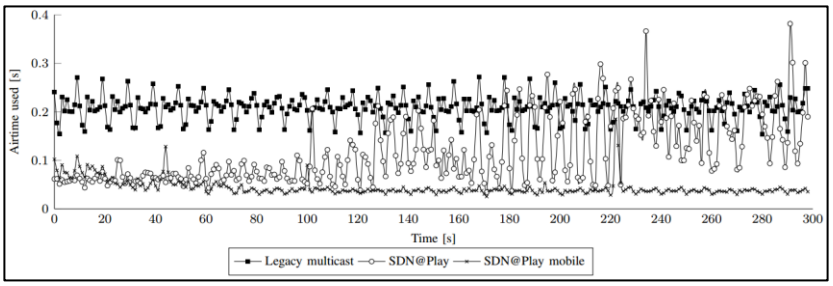


Figure 4-25 Channel utilization over time of the AP1 for the multicast video transmission at 1.2 Mbps

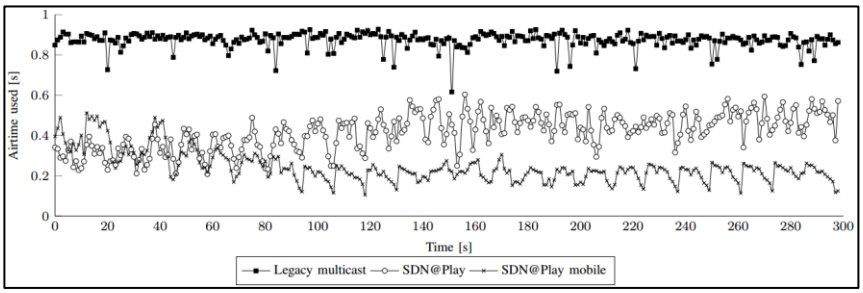


Figure 4-26 Channel utilization over time of the AP1 for the multicast video transmission at 6.2 Mbps

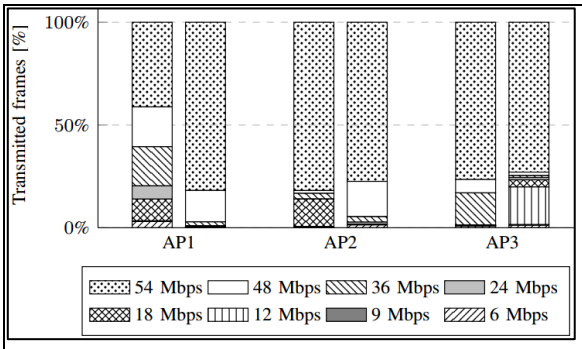


Figure 4-27 Distribution of rates in SDN@Play & SDN@Play Mobile per AP at 1.2 Mbps

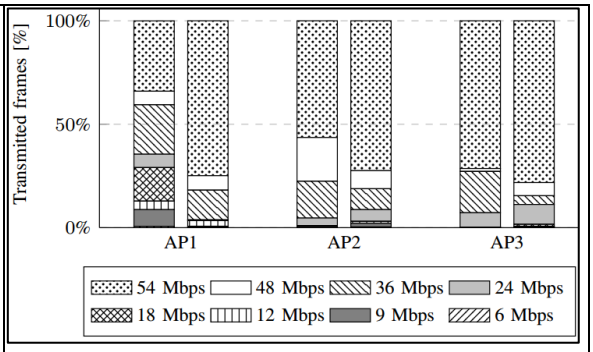


Figure 4-28 Distribution of the rates in SDN@Play and SDN@PlayMobile per each AP at 6.2 Mbps

4.2.3.7 Conclusion

A novel multicast rate adaptation and mobility management scheme for 802.11–based WLANs was presented. The proposed scheme uses an SDN approach where the global network view available at a logical centralized controller is exploited in order to coordinate the operations of different APs. The proposed scheme, named SDN@Play Mobile, jointly optimize the multicast rate selection at the various AP and the multicast receptor association with the overall goal of reducing the network–wide radio resource utilization while maintaining the expected quality of experience. SDN@Play Mobile has been implemented and evaluated over a real–word testbed using the 5G–EmPOWER platform. Experimental measurements show that SDN@Play Mobile can deliver a significant improvement in terms of channel utilization compared to the Legacy multicast in 802.11 while maintaining full backward compatibility with standard 802.11 wireless terminals.

4.3 Relaying & D2D

This section collects studies on D2D communications and relaying. The first sub-section presents a MAC model of D2D communications in autonomous mode (i.e. out-of-coverage and/or without network help) in the context of LTE-A Pro. The following two sections both deals with D2D communications. The first one uses out-of-coverage D2D communications and the concept of User Equipment Relay to implement coverage extension. The COHERENT architecture is used for dealing with the multiple cases of possible topologies and to flexibly instantiate protocol stacks according to the needs of the applications. The third subsection presents how the COHERENT architecture can be used for enhancing end-to-end DL throughput in a multi-cell environment, by

controlling two-hop D2D relaying. The last two subsections deal more with the relaying concept. The fourth subsection presents the moving relaying concept in the context of air-to-ground communications. Finally, the moving cell concept is investigated by the last contribution of this section. Here, existing LTE features are combined together in an innovative way to create a wireless mesh. The COHERENT controller is used for solving resource allocations issues between infrastructure links and access links in the mesh.

4.3.1 Modelling of SideLink Communications in LTE-A

4.3.1.1 Motivation and goals

In this section we derive models for Device-to-Device (D2D) communications for 3GPP Rel 13 and 14, i.e. LTE-Advanced (LTE-A) Pro. D2D communications are called SideLink (SL) communication in the standard and they were initially specified also for Private Mobile Radio (PMR) applications. Both PHYsical (PHY) and Medium Access Control (MAC) behaviours are studied. Concerning MAC, a model of the probability of successful access to the medium is derived and an abstracted metric is proposed, namely an achievable throughput which is able to summarize the impact on data rate of MAC and PHY features of LTE-A Pro specification.

This work is related to use case 3.CE described in COHERENT D2.1 [2]: "Coverage extension and support of out-of-coverage communications". Autonomous D2D communications are treated here since they can be used out of coverage. However, this model can be useful to D2D communications in coverage, if the network decides to leave the users autonomous. Moreover, the work is really intended to model the D2D communications of the LTE-A Pro standard *as it is* (without improvements), providing tools necessary for a detailed assessment of its performance.

Since it was not possible to include all the details in this deliverable, the interested reader can refer to a COHERENT technical report [9], or to conference publication [21].

4.3.1.2 Requirements

Table 4-17 presents the COHERENT system requirements in [2] related to this study. This contribution in fact provides a way to abstract MAC behaviour of SL communications in LTE-A Pro (R#1) and hence it may help monitoring and control, in particular with respect to the optimization of spectral efficiency (R#62). It also addresses resilience and high availability (R#9, R#51, R#55, R#69) by modelling a technique of direct communication which is important for Public Safety (PS) services. Finally, since it deals with the modelling of the LTE-A Pro SL communications in autonomous mode as they are specified in the standard, it contributes to the inclusion of legacy 4G and evolved 4G systems into the COHERENT framework.

Table 4-17: Requirements from COHERENT D2.1 [2] addressed by modelling of D2D communications in LTE-A Pro

R#1	COHERENT shall ensure the necessary physical and MAC layer abstraction
R#9	COHERENT shall support PPDR services
R#11	COHERENT shall be able to manage future broadband PMR systems
R#12	COHERENT shall be able to manage legacy 4G systems and solutions
R#51	COHERENT shall be able to provide minimal services like voice, localization, transmission of pictures with high reliability for PPDR applications (typically after a natural disaster).
R#55	COHERENT shall support out-of-coverage communications for PMR services
R#62	COHERENT shall be able to monitor and control network resources such that spectral efficiency is increased
R#69	COHERENT shall have a fallback mode of operation in case of natural disaster or when the fixed infrastructure is out-of-order

4.3.1.3 Modelling of SL PHY

In this subsection we deal only with understanding the PHY layer behavior of SL data channel, named Physical Sidelink Shared Channel (PSSCH). We also studied a PHY layer model of the SL control channel, named Physical Sidelink Control Channel (PSCCH), and of PSSCH which is not included here for space issues. For detailed descriptions please refer to [9].

For clarity's sake a brief summary of the operation of SL communications in LTE-A Pro is recalled. Much more details are provided in COHERENT D3.1 [4]. At PHY and MAC layer, D2D communications in LTE-A Pro are organized inside one (or multiple) set of resources called Resource Pool (RP). In the frequency domain, a RP is defined by N_{RB} Resource Blocks (RB) divided in two sub-bands of equal width. A RP is temporally described by a Sidelink Control (SC) period (L_{SC} subframes long), divided into a first part dedicated to the transmission of control information (PSCCH) and a second part for data transmission (PSSCH). The SC period is repeated, thus providing a sort of new radio frame designed for SL communications.

In autonomous mode, the access is random and contention-based. First, the UE independently chooses where to send the SL Control Information (SCI) over the PSCCH (control part). The RB choice is done in a uniform random way, according to predefined rules (see [73] for more details). Then, for data transmission in the PSSCH, the UE randomly selects the allocation too, which is encoded in the SCI, valid for one SC period, and which consists in the choice of L_{CRB} contiguous RBs as well as the scheduled sub-frames in the time domain.

PSSCH uses a fixed pre-configured Modulation and Coding Scheme (MCS), which determines the modulation and the code rate used by LTE turbo encoder. When a packet is ready for transmission, HARQ protocol without feedback is applied, i.e. the packet is always (re)transmitted four times without waiting for feedback. In detail, the packet is coded and four Redundancy Versions (RVs) are created. Then, RV 0, 2, 3, and 1 are sent in consecutive scheduled sub-frames. If the UE buffer is not empty, another packet is processed and sent in the next four consecutive scheduled sub-frames, according to the same allocation information, until the end the SC period is reached.

This procedure has been designed to provide a robust and redundant communication mode, since, due to the completely random choice of the allocation resources, a packet loss may happen because of: 1) collisions at the receiver side, if two or more transmitters choose the same resources; 2) Half-Duplexing (HD) constraints, i.e. a UE acting as a transmitter, when scheduled for SL transmission, cannot activate its receiver for listening to other transmitting UEs.

Here we want to understand the impact of HD constraints on PHY layer performance. It is then important to briefly describe how HARQ RVs are generated by the rate matching algorithm (a tutorial can be found in [16]). The coded bits at the output of the LTE turbo code with rate 1/3 are composed by systematic bits (equal to information bits), parity 1 and parity 2 bits. These three groups of bits are separately interleaved and then multiplexed in a so-called Virtual Circular Buffer (VCB), which can be visualized as a table with a fixed number of columns (namely 96) and with a variable number of rows depending on the input packet size. Systematic bits occupy the first third of the columns, while multiplexed parity 1 and parity 2 bits occupy the remaining columns.

The rate matching algorithm generates RV i by starting to read the VCB column-wise at a predetermined column c_i. If the end of the virtual circular buffer is reached, it continues to read the VCB starting from the beginning, until the number of bits required to fill the resources allocated for the transmission is reached. Since systematic and parity bits are not mixed together in the VCB, their distribution inside a RV is not homogeneous and depends on the offset c_i. For code rates $R_c >$ 1/2, RV 0 contains most of systematic bits. Then, the percentage of systematic bits decreases for RV 3, 1 and 2. Now, besides a sufficiently high SNR, turbo decoding needs also a minimum amount of systematic bits for starting to converge towards the correct sent codeword [20]. Below a given percentage of systematic bits, it is impossible to decode the packet even in absence of noise. This effect is shown in Figure 4-29, where we plot the E_s/N_0 loss in dB of RV 1, 2, 3 with respect

to RV 0 at a target Packet Error Rate (PER) of 0.1 for a subset of LTE MCSs, in an additive white Gaussian noise channel. Losses increase with increasing code rate R_c of RV 0. Filled markers represent non-self-decodable RVs. Starting from $R_c \approx 0.67$ RV 2 is no longer self-decodable, then from $R_c \approx 0.75$ RV 1 as well is not self-decodable and finally RV 3 at very high R_c.

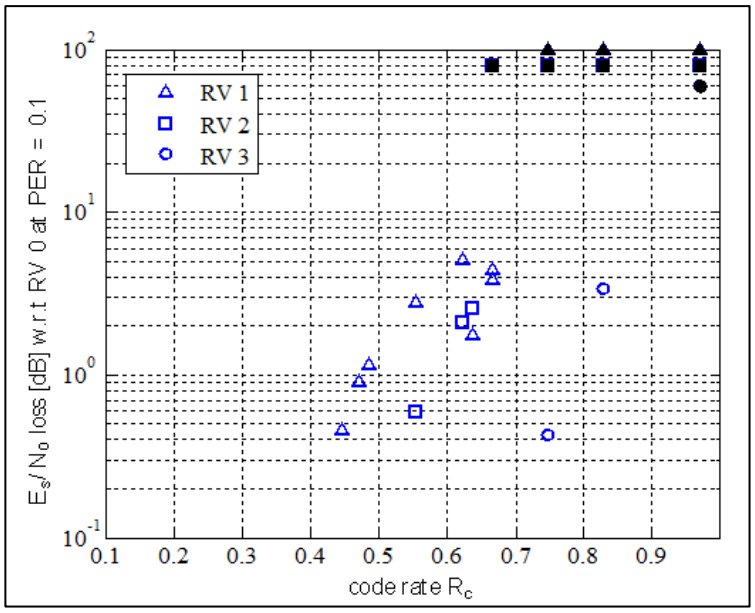


Figure 4-29: E_s/N_0 loss in dB of RV 1, 2, 3 with respect to RV 0 at a target PER of 0.1 for some MCS; filled markers represent non-self-decodable RVs.

Many MAC models consider a successful packet reception when at least one of its (re)transmissions (i.e. RVs) is not lost or collided. However, if losses are due to half-duplexing constrains, soft bits cannot be accumulated by the UE, since it is in transmit state. In case of a unique received RV, successful decoding even at very high SNR happens only when the RV is self-decodable. Hence, it is important to include this effect in the evaluation of the MAC access probability. Finally, this strong dependency of performance on the effectively received RVs implies that low complexity PER prediction methods (also generically called link-to-system mappings) like the one proposed in [17] are not well suited for the SL case. We suggest to use what [17] calls a brute force method, i.e. using a different PER curve for each combination of RVs, combined with a traditional mutual information model, i.e. a mutual information effective SINR metrics (MIESM), like those in [17], [18], [19]. For further details, please refer to [9].

4.3.1.4 Modelling of SL MAC

In this subsection we deal only with modeling the MAC layer behavior of PSSCH. We also studied a MAC layer model of PSCCH, building on top of results in [73]. For sake of conciseness, only a summary of results for PSSCH is reported. For detailed results please refer to [9].

LTE-A Pro uses Time Resource Pattern (TRP) of Transmission, a short bitmap, for indicating scheduled sub-frames in the data part of the SC period. TRP has length L_{TRP} equal to 6, 7, or 8 bits, depending on the LTE-A Pro FDD/TDD frame configuration used for SL [38]. The positions of TRP bits equal to 1 indicate the position of scheduled sub-frames inside the PSSCH part of the SC period. The length L_{PSSCH} of the PSSCH can be much longer than L_{TRP}, hence TRP is repeated periodically up to the end of the resource pool. The weight w of the TRP, i.e. the number of its bits equal to 1, may take values in the range 1, …, L_{TRP}.

In the frequency domain, SCI gives only one RB allocation per SC period. Frequency hopping (FH) can be used; otherwise the initial fixed RB allocation is kept. FH type 1 is an explicitly signaled hopping pattern, while FH type 2 is a pseudo-random FH pattern hopping in $N_{sb} = 1, 2, 4$ separated sub-bands of equal size. However, a detailed analysis shows that for all FH types, inside one RP, FH patterns are deterministic, depend on the RB allocation, and do not depend on user identity. Hence, if a collision happens in the SCI RB allocation, it will be repeated over the whole hopping pattern. Only if the selected TRPs are different, the non-collided sub-frames will allow discriminating the signals.

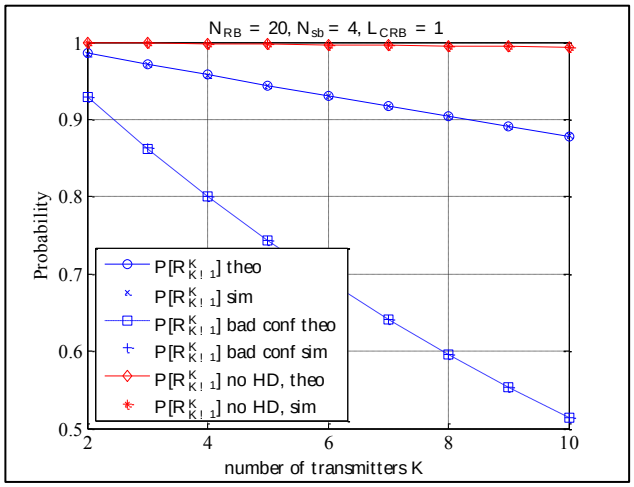


Figure 4-30: Probability $\boldsymbol{P}\left[\mathcal{R}_{K-1}^{K}\right]$ for a variable number of users accessing PSSCH with 20 RBs divided in 4 sub-bands, with packets of 1 RBs

Skipping a lengthy derivation (see [9], [21], and [73]) we present, for the PSSCH, the probability $P[\mathcal{R}_k^K]$ that k users among the K transmitters correctly receive the packet of one transmitter, under no capture effect (i.e. interference is strong when collisions happen)

$$P\left[\mathcal{R}_k^K\right] = [1-(1-P[\mathcal{X}])^{K-1}]\delta(k) + \binom{K-1}{k}\left(1-\frac{N_{bad}}{N_{TRP}}\right)^k \left(\frac{N_{bad}}{N_{TRP}} - P[\mathcal{X}]\right)^{K-k-1} \qquad (4\text{-}4)$$

where $N_{TRP} = \binom{L_{TRP}}{w}$ is the number of different TRP patterns with length L_{TRP} and weight w, N_{bad} is the number of bad decoding configurations due either to perfect collision or non-self-decodable RVs, and $P[\mathcal{X}]$ is the perfect collision probability, i.e. the probability that two users choose exactly the same RBs and TRP. For an analytical formula of $P[\mathcal{X}]$ and N_{bad}, please refer to [9], [21]. When $k = K - 1$, then $P[\mathcal{R}_{K-1}^K] = \left(1-\frac{N_{bad}}{N_{TRP}}\right)^{K-1}$. This probability will be used in the following section, since it provides an expression for successful access for all the users accessing the channel. An interesting remark is that $P[\mathcal{R}_{K-1}^K]$ is independent of the number of resource blocks. This is due to the fact that half-duplexing constraints are the dominant effect and, in LTE-A Pro design, they are controlled only by TRP. In typical SL configuration over an FDD frame, $N_{TRP} = \binom{8}{4} = 70$, which in fact is not a high number of possibilities for avoiding HD constraints. The presence of non-self-decodable RVs starting from code rate equal to 2/3 is worsening the situation even more. The impact of HD constraints and of bad decoding configurations (one non-self-decodable RV leads to $N_{bad} = 5$ in LTE-A Pro design) on $P[\mathcal{R}_{K-1}^K]$ with respect to a case without such problems is illustrated with a numerical example in Figure 4-30.

4.3.1.5 Proposed abstraction at MAC level

We propose a total weighted achievable throughput metric, which represents the maximum aggregated throughput of the SL channel

$$\eta_K = K N_{p,av} b_{MCS} P_s^{SL} / L_{SC} \quad (4\text{-}5)$$

Where K is the number of active users, b_{MCS} is the number of information bits supported by the chosen MCS, L_{SC} is the length of one SC period in seconds, $N_{p,av}$ is the average number of packets that can be sent in the SC period, and P_s^{SL} is the probability of successful access for the global SL channel, i.e. for the control and data channels. The previous equation holds true because in the standard all the users must use the same MCS, and we suppose the same traffic type for all the UEs. $N_{p,av}$ is a random variable which depend on the structure of PSCCH and PSSCH (see [9] or [21] for further details). P_s^{SL} is defined as follows:

$$P_s^{SL} = P_s^{PSCCH} P_s^{PSSCH} = P^{PSCCH}[\mathcal{R}_{K-1}^K] P^{PSSCH}[\mathcal{R}_{K-1}^K]. \quad (4\text{-}6)$$

and it is a function of K and of the parameters of both the PSCCH and PSSCH, such as the number of frames dedicated to PSCCH L_{PSCCH} and to PSSCH L_{PSSCH}, N_{RB}, N_{TRP}, L_{CRB}, N_{sb}, N_{bad}, etc.. Finally, by setting a target minimum success probability $P_{s,target}^{SL}$, it is possible to impose a quality constraint on the access probability $P_s^{SL} > P_{s,target}^{SL}$, which impacts the maximum number of simultaneous users K accessing to the medium. Thus, η_K takes into account:

a global guarantee of success of the overall SL communication at MAC level $P_{s,target}^{SL}$

structural constraints of the SL definition in LTE-A Pro (through L_{PSCCH}, L_{SC}, TRP parameters, FH parameters, etc.) the MCS throughput as well as the presence of possible bad decoding configurations.

SL parameters are input to the abstraction and can be hidden by the value of the weighted achievable throughput. At the same time, depending on the control function and implementation constraints, certain parameters can be controlled by the function.

As an example, in the following, we may want to maximize the weighted throughput with a constant number of available RBs, and total length of the SC period (which are imposed, say, by a latency requirement), but being able to tune the length of the control part and the number of users.

Thin curves in Figure 4-31 represent η_K as a function of the length of control part L_{PSCCH}, for different K and a LTE-A Pro setting given by $L_{TRP} = 8$, $w = 4$. The RP parameters are $N_{RB} = 20$, $N_{sb} = 4$, $L_{SC} = 40$, and packet size $L_{CRB} = 1$ RB, which fits well to PS voice application. Two MCSs are considered: MCS 1 carries $b_1 = 224$ information bits with $R_c = 0.47$ and has no bad decoding configuration; MCS 2 carries $b_2 = 328$ bits with $R_c = 0.67$ but RV 2 is not self-decodable, which corresponds to $N_{bad} = 5$ in our model.

The thin dashed and solid curves correspond to η_K respectively for MCS 1 and MCS 2; different colors represent different K. When $K = 1$, no collision is present, hence the optimal length of the control part is 2 sub-frames (i.e. its minimum value in LTE-A Pro). In this case MCS 2 provides higher achievable throughput. When $K > 1$, collisions and HD issues are present. If the control part is too short, the throughput falls due to low $P^{PSCCH}[\mathcal{R}_{K-1}^K]$. Then a maximum is achieved and the curves start decreasing, since a too long control part penalizes the number of packets that can be

sent in one SC period. For fixed K, curves of MCS 2 are always higher than the corresponding curves of MCS 1. However, these thin curves are drawn without setting a constraint on the successful access probability.

If the system requires a minimal probability of successful access, then conclusions change and are represented by the thick curves (dashed for MCS 1 and solid for MCS 2) in Figure 4-31. All the throughputs points below or on the thick curves satisfy the access quality constraint $P_s^{SL} > P_{s,target}^{SL}$ = 90%. The maximal K supporting the quality constraint changes with the control part length. Moreover, MCS 2 has bad decoding configurations, hence the satisfaction of the quality constraint is much harder (see also Figure 4-30). For MCS 2, only curves for K = 1, 2 are drawn because for K > 2, the constraint $P_s^{SL} > P_{s,target}^{SL}$ = 90% is never satisfied. In summary, MCS 1 provides a much larger set of configurations than MCS 2 satisfying the required minimum successful access probability. Hence, choosing MCS 2 becomes suboptimal if we want to maximize η_K for medium lengths of the control part, or if K > 2. However, MCS 2 can be the right choice if the number of concurrent transmitters K is always very limited.

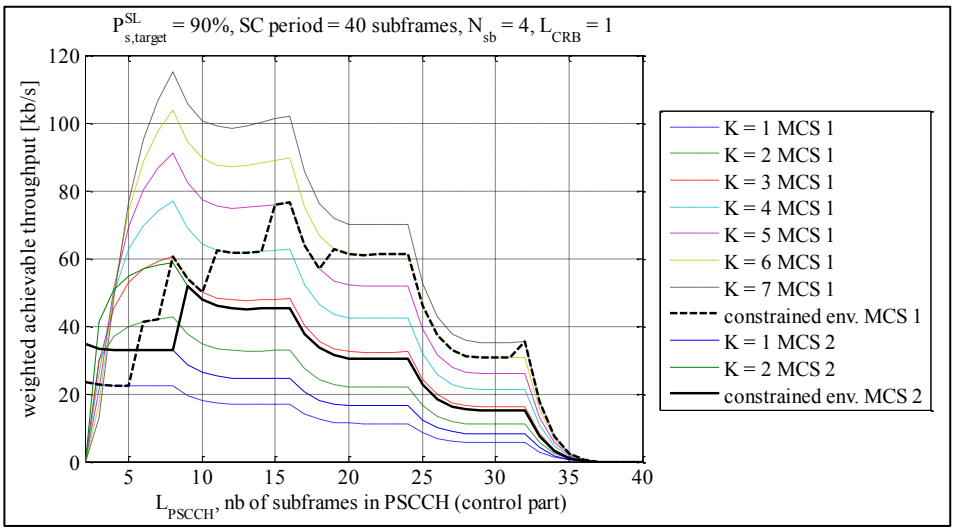


Figure 4-31: Total weighted achievable throughput with a target probability of successful access of 90%; points below or on thick curves satisfy the constraint

4.3.1.6 Conclusion

In this contribution, we provide an analysis of SL PHY and MAC layer behaviour and a MAC layer model for data channels of SL at PHY level. Further results on PHY layer models or SL control channel are responded in the COHERENT Technical Report [9]. For the data channel, we show that the current LTE rate matching algorithm is not optimal under strong half-duplexing constraints, since for high code rates there exist non-self-decodable Redundancy Versions, i.e. retransmissions that cannot be decoded alone. This fact suggests that it is better to resort to brute force link-to-system mappings, since the traditional more compact methods based on a reduced set of curves may give very optimistic predictions. It is very important to take into consideration these non-self-decodable RVs also when assessing MAC level throughput performance. For SL data channel, analytical formulas are derived of the probability that any subset of the transmitters correctly receive the transmission of one given user. The effect of half-duplexing constraints and bad decoding configurations is included. Finally, the global successful access probability for SL communications is derived, which adequately characterizes the performance of the current LTE-A Pro SL scheme in autonomous mode.

Building on top of the previous finding, a weighted achievable throughput has been defined taking into account the access probability, structural constraints of the SL data channel as defined in LTE-A Pro, and the used modulation and decoding constraints. This metric precisely characterizes realistic throughputs attainable by LTE-A Pro direct communication in autonomous mode as a function of the system setting. It is a good abstraction for all control applications dealing with throughput in SL. The analysis of the formulas and of simulation results on metric values shows that LTE-A Pro SL communications are reserved to a very small number of concurring transmitters, if a high probability of successful access is imposed by the applications. This fact must be carefully considered when using SL communications for public safety services. Another added value of our analysis is that it gives hints about possible improvements in LTE-A Pro SL communications. Simple ideas for future investigations range from parameter tuning, e.g. increasing the TRP length for decreasing losses due to HD constraints or defining FH per user, up to PHY and MAC modifications, e.g. defining a better rate matching algorithm for SL or proposing new access procedures.

4.3.2 Coverage extension through D2D for LTE and evolutions

4.3.2.1 Motivation and goals

This study is aligned with latest 3GPP advancements on UE-Relaying and network control of remote UEs. Recently, 3GPP started two Study Items on "Remote UE access via relay UE" (REAR) and on architecture enhancements to "Proximity Services (ProSe) UE-to-Network Relay" (FS_REAR) at SA1 level (for Release-14) and SA2 level (for Release-15) respectively. In the proposed study items, it is said that remote UEs (UEs-out-of-coverage) can employ D2D or other non-3GPP short-range communication technologies to access to the network via the UE-Relay (UE-R hereafter and also named by 3GPP "UE-to-Network Relay"). For the complete set of goals please refer to the updated COHERENT M3.2 [5] (see SP-160961, SP-160834, S2-166212 and TR 23.733 [36]). We recall few goals of these studies and in particular those which are completely in-line with the COHERENT approach:

- Procedures and solutions to have network control of Remote UEs through a UE-R.
- Flexible architecture allowing any UE to become a UE-Relay, and which consequently needs single-hop & multi-hop relaying.
- Dynamic reconfigurations of the interfaces for L2 and L3 layers (see e.g. TR 36.746).
- Multi-connectivity, improved mobility and handover control.
- Congestion control and traffic/throughput management for Remote UEs and UE-R.
- Voice services for Mission Critical Push to Talk (MCPTT) in SA6.
- Energy efficiency for Remote UEs.
- Agnostic control and generic interfaces, as proposed by COHERENT.

4.3.2.2 Requirements

Section 4.3.2.1 shows the interest of 3GPP for studying how to extend the control of D2D links both in-coverage and for coverage extension through a UE-R. COHERENT control plane can be therefore extended to control remote UEs (i.e. UEs out-of-coverage), and different situations have been further imagined with respect to this goal:

- Topology for Single-Hop[4] communication (with one UE-Relay).
- Topology for Multi-Hop communication (with multiple UE-Relays).

[4] Please note that in this section, the eNodeB-UE link is not considered as a hop, hops are counted starting from the first UE trasmitter to the last UE receiver.

- Topology for Broadcast-type communication (for virtual Radio Processing (vRP) or UE-R, see Sec. 0).

As a matter of fact, an important requirement is the ability of the system to adapt to topology changes. The problem is how to reconfigure the system in order to reduce e.g. its latency, and potentially other QoS parameters such as throughput, BER or PER, in order to support use cases on D2D links and in particular coverage extension.

Moreover, each of these topologies may require an adaptive configuration of the protocol stack as a function of the real-time scenario, which will be further discussed in the next subsections. . In some situations, a topology might change (as a result of the environment change: new user connecting to the network, or evolution in time of the interference, BER or PER, or evolution of required or real-time throughput) which will require an adaptation of the configuration, and for this reason we sometimes refer to "adaptive" configuration, or "flexible" and "real-time" configuration. For explicit COHERENT topologies please refer to M3.2 [5]. In the case of multi-hop topology parameters such as latency and others have to be taken into account when designing access, mobility and control strategies. Other topologies can be envisaged for broadcast mode, with two UEs receiving the same data flow, or two remote UEs receiving the same data flow through one UE-R (one single broadcast bearer) at the same time. These topologies will require some other different configurations, which have been already presented to some extent in COHERENT D2.2 [3]. For this reason, this contribution will mainly focus on configurations required for single-hop and multi-hop communications, in the next subsections.

4.3.2.3 Architectural and control aspects

The COHERENT framework applied to network coverage extension and mobility management is schematically represented in Figure 4-32, where different functional blocks have been described for RTC/C3 and for network and equipment entities such as UE-R or eNB (with both R-TP and vRP, see Sec. 0). Three main components of the RTC/C3 are represented as follows:

- A network graph block which can be distributed between C3 and RTC (RTC has a view on the short-term processes while C3 on the long-term processes);
- A decision block which can be on C3 or RTC, and which is responsible of the decisions for dynamic configuration of network entities and user equipment, such as protocol stack re-configuration at different layers with respect to topology and measurements provided by the network graphs;
- A configuration block, at RTC level, which is responsible for configuration of protocol stack, measurement configurations, measurement reports configurations, coverage extension and potentially for mobility management of a remote UE.

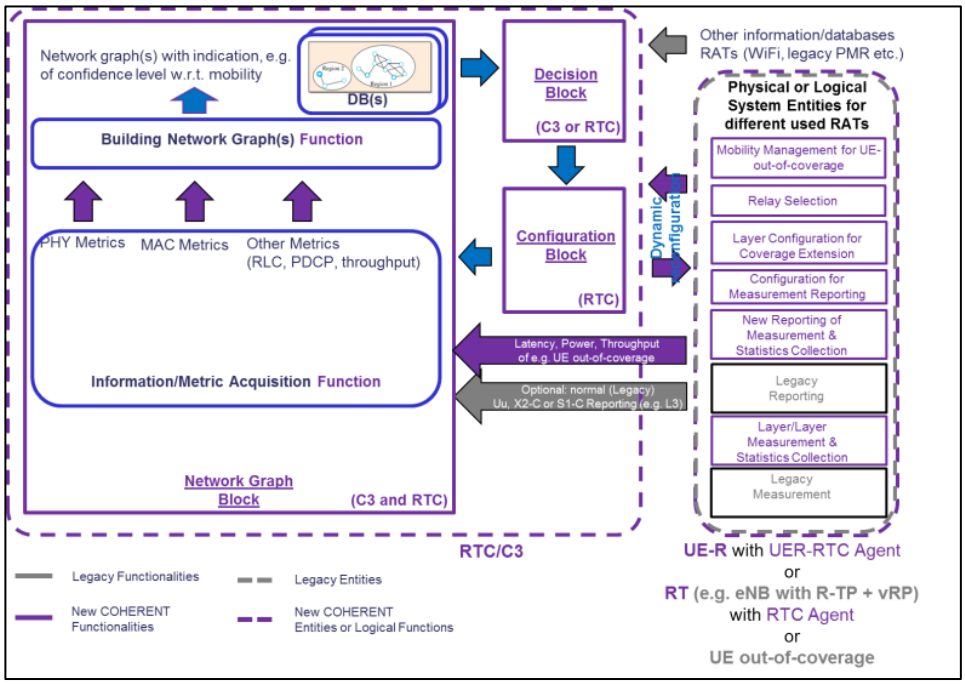


Figure 4-32: COHERENT Architecture for Coverage Extension and Mobility Management

As introduced in Sec. 2.2, at UE-R level, RT level and Remote UE level (if required) there are 1) legacy functionalities such as legacy configuration, legacy measurement and legacy reporting, and 2) improved functionalities added by COHERENT framework which include layer per layer measurement and statistics collection, new reporting of measurement and statistics collection under the control of a new manager of measurement reporting, protocol stack (sub-)layer configuration for measurement reporting, relay selection functions and mobility management for a Remote UE (e.g. UE-out-of-coverage). All these functions can be implemented with the help of an RTC Agent (see Sec. 2.2); for example in Figure 4-32, on the right hand side we have shown UE-R and eNB (R-TP + vRP) entities each with UER-RTC and RTC functionalities respectively.

In this proposed COHERENT architecture, the RTC and C3 blocks are used to dynamically reconfigure the physical or logical system entities with respect to different RATs. RTC/C3 may therefore deal with mobility management of remote UEs, to perform relay selection, Handover (HO) for remote UEs or UE-R, and configuration for coverage extension. The information needed for these control applications can use legacy reporting similar to the one used for Uu, X2-C or S1-C in case of LTE RAT, or use new reporting.

The system flexibility given by the use of RTC/C3 is allowing not only to adapt specific control applications like, say, MAC scheduling, but also to decide what protocols to be used and in which conditions. For example, RTC/C3 may decide to add an extra RRC UE protocol just for remote UE purposes, which will be encapsulated in Packet Data Convergence Protocol (PDCP) protocol of a UE-Relay, or to remove PDCP protocol for a UE-Relay and use only the PDCP protocol of a remote UE, etc. This capability will have an important impact on the protocol stack performance as a result of processing latencies and potential service interruptions. However, a way to implement this protocol stack re-configurability can be obtained through virtualization [14]. The impact on the protocol stack reconfiguration time can be alleviated by having multiple active vRP instances (one for each protocol stack configuration) on the same physical entity, which can be reproduced, saved, multiplied, erased or modified with respect to the real-time requirements.

4.3.2.4 Network graph and abstraction aspects

Since our proposal of protocol stack re-configurability work on layer (or sub-layer[5]) basis, it is important to have measurements and network states on a layer basis too. Hence, the first main core idea expressed in this section is that network graphs should not be seen as a simple collection of data, but as a collection of data which is obtained on layer per layer basis. Moreover, for different network topologies one may have a complete distinct set of network graphs. At each level (layer/protocol and entity), RTC/C3 may extract information about different measurements, connectivity information, throughput, latency and even security, in order to better identify when and how performance can be optimized. In this way, RTC/C3 has complete information about the network and may decide how to optimize it.

All network graphs could be further divided in two separate sets (as on Figure 4-33): 1) Requirements, dictated by services; 2) Achievements, with achieved real-time values. In practice, the real-time metric is not necessary instantaneous, and can be modelled over a short duration of time to obtain mean values, statistical values or probability distribution. These two network graph visions are used by the control applications to determine or jointly-optimize, for example, the configuration of a network in order better to respect the system constraints. On the other hand, it is easy to see that, depending on the system cost or the service requirement, certain constraints (defined layer by layer) are contradictory: for example, it is not always possible to increase the useful flow rate without increasing the latency or decreasing the robustness. This is why we need an intelligent algorithm to independently manage per-layer configurations, depending on the system requirements. The exact description of Figure 4-33 and some network graph definitions used in this section are given in M3.2 [5], comprising connectivity graphs, latency graph and more generalized definition of network graphs.

4.3.2.5 Solution description

The goal of the solution is to offer a transparent connection to a remote UE (i.e. no modification of the protocol stack is required from the remote UE perspective), while improving system performance with respect to throughput, interference reduction, handover and latency (in multiple forms such as protocol latency, latency resulted from interruptions during handover, or latency resulted from de-prioritization of the bearer with respect to other less-important transmissions).

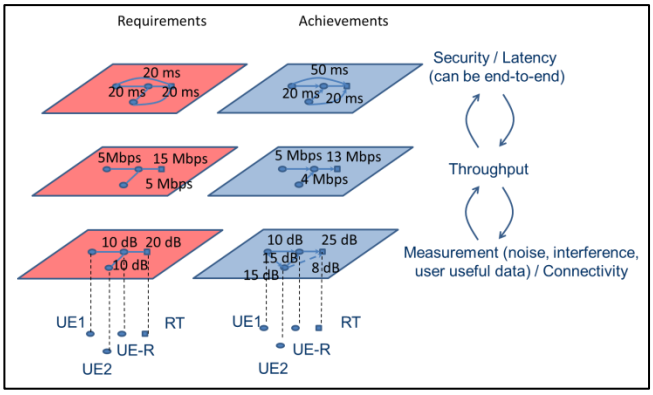


Figure 4-33: Example of Network Graphs Information at RTC/C3 level

[5] For instance, MAC, RLC, PDPC are sublayers of L2 protocol stack of LTE.

The proposed algorithm determines the configuration to be applied, selecting it from a feasible set of solutions, depending on the topology of the network. The proposed algorithm chooses what type of protocol stack configuration has to be applied to the existent network topology. “Single-hop” and “multi-hop” control plane configurations are represented respectively in Figure 4-34 and in Figure 4-35. For user plane configurations please see M3.2. Note that the algorithm is oriented to LTE protocol stack, as far as RT-UE and RT-core communications are concerned. Moreover, with respect to all the figures below, the interface PHY/CPRI (PHYsical layer/Common Public Radio Interface) can be replaced for example by any other generic interface such as SCTP/IP/L2/L1 or MAC Ethernet/PHY Ethernet (in case of the control plane representation). The protocol stack between R-TP and vRP is not therefore fixed, and the API can allow for example the deployment of MAC functionalities on R-TP level, depending on the functional split between R-TP and vRP.

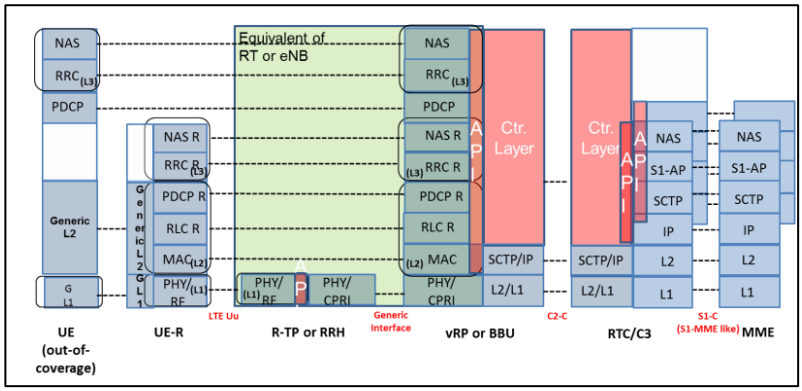


Figure 4-34: Single-Hop Relaying configuration (Control Plane)

The configuration option represented by “single-hop relaying” in Figure 4-34 for Control Plane describes a PDCP UE layer encapsulated into the PDCP UE-R layer. The UE-R relays the information towards the UE without being able to decode it. The R-TP implements lower layers while upper layers are implemented at vRP level. Between UE-R and UE there are some generic L2 and L1 layers and the data flow for UE-R is forwarded to the upper layers while the data flow for UE out-of-coverage is forwarded to generic L2 and L1 layers. An equivalent of the Radio Link Control (RLC) layer can be considered on the interface between UE and UE-R (i.e. inside the generic L2). The remote UE (out-of-coverage) is hence controlled by the vRP, which performs functions such as RRC and NAS. RTC/C3 recovers information from Mobility Management Entity (MME) and, through the new COHERENT control layer, performs the remote measurement configuration and reporting required for the network graphs. The configuration and reporting operations can be distributed between the RTC/C3 and vRP but RTC/C3 is responsible for the final decision/control operation. Multiple applications may run on UE side (out-of-coverage) - parallel applications at IP level in the figures corresponding to user plane from M3.2. At the same time, a remote UE relayed by UE-R may be attached to other MMEs (or core networks), which explains the parallel protocol stacks in the RTC/C3 represented on the previous figures.

At least two kinds of configurations are envisaged:

- Configuration of parameter measurement (layer by layer) for the network graphs located in RTC/C3.
- Configuration of layers (e.g. adding/removing layers), for example adding or removing PDCP layer for UE-R and/for Remote-UE.

The configuration option represented by "multi-hop relaying" in Figure 4-35 describes a situation where the UE-R performs the control of the remote UE. In DL, UE-R may decrypt the information addressed to the remote UE and can possibly re-encrypt it before relaying it. In UL too, the UE-R is able to decrypt the information sent by the out-of-coverage UE before relaying it to the network. This may apply to a use case where a policeman is relaying information to a fireman while being able to listen to the conversation. UE information can also be encrypted at application level and a UE-R relaying function may help to map the packets toward the radio bearer(s) to be sent to the remote UE. In this case UE-R cannot decode the UE information before relaying it if not sharing the keys of the remote UE.

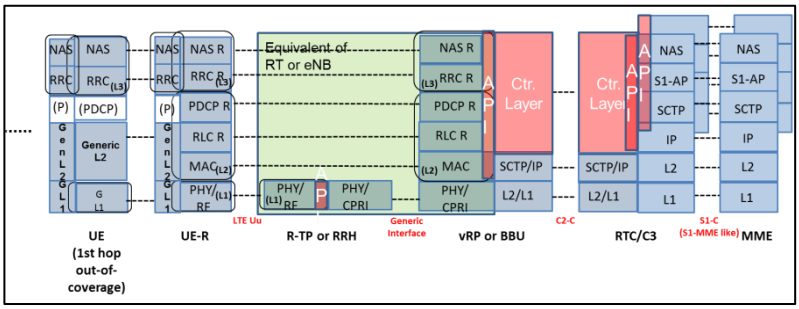


Figure 4-35: Multi-Hop Relaying configuration (Control Plane)

The interface between R-TP and UE-R and the interface between UE-R and remote UE at L2/L1 layers are similar to the ones described in Figure 4-34. An equivalent of the RLC layer can be considered on the interface between UE and UE-R, e.g. as a part of the generic L2 layer. A faster and more responsive control (similar to RRC or NAS) may be directly provided by the UE-R close to the remote UE. The remote UE is no longer controlled by the vRP and therefore reducing the control latency introduced by multi-hop communications.

RTC/C3 still recovers information from MME, UE, UE-R, S-GW and P-GW and, through the new control layer, it performs the decisions and the (remote) configuration of measurements, whose operation can be distributed on different entities as in the "single hop" case. In the case of RTC/C3 collecting multi-hop measurements, a relaying function may be involved in relaying measurement information collected at the remote UE side. The multi-hop configuration can be applied to multiple series of UEs connected to UE-R. In this case, intermediate UEs can become UE-Rs from the point of view of the relayed UEs, as represented in Figure 4-35.

Figure 4-36 represents an example of logic flow where connectivity graphs, measurement graphs and latency graphs are continuously updated by RTC or C3. All the network graphs (NGs) can be contained in a database used for saving, updating and requesting information. Moreover, the network graph information can be accessed using Network Information Functions (NIFs) [3]. In some situations, the control can be distributed between RTC and C3, for example for inter-cell interference coordination over a large area. However, in general it is preferred to have the configuration and control applications as close as possible to the physical equipment, and therefore on RTC side, for real-time tasks. This representation is in line with what said for Figure 4-32, where it has been claimed that for UE-R applications the configuration should be mainly on RTC, the decision on RTC or C3, and the network graph database on RTC and C3.

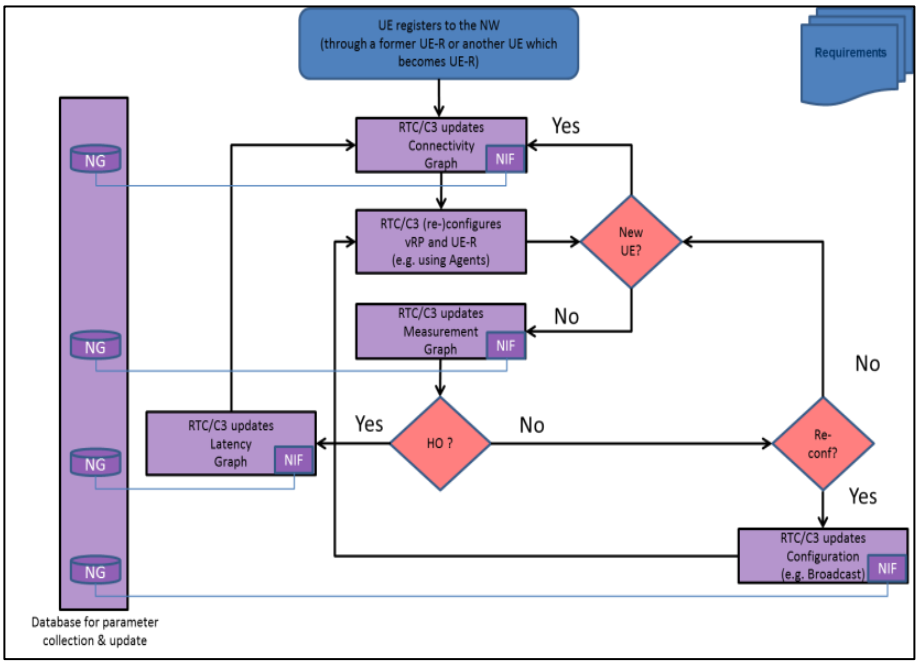


Figure 4-36: Example of Potential Configuration Algorithm at RTC/C3 level

As presented in Figure 4-36, the algorithm starts with a UE registering to the network. Then RTC/C3 updates the network connectivity graph with new information and then configures the UE-R with new parameters/protocols and the vRP using configuration agents UER-RTC and RTC. Then, the algorithm checks if a new UE (potentially with other service requirements) has registered to the network or not. If a new UE (or UE-R) has registered, then the RTC/C3 updates the connectivity graph. If not, then RTC/C3 updates the measurement graph and then checks if a HO is required. If HO is not required, then the RTC/C3 can check for example if further optimisation of the network is possible. For instance, if in the out-of-coverage group only group communication is used, it can choose a broadcast mode, where the same physical resource is used to send common information to multiple users at the same time. If HO is required, RTC/C3 updates latency graph with e.g. information related to maximum interruption time during HO and then connectivity graph is updated with new information resulted from HO operation, since the topology of the network may change. This contribution is complementary with the other contributions in this deliverable in the sense that it describes a general framework rather than a precise parameter optimisation algorithm.

Depending on the given radio access topology, RTC/C3 updates the network graph with new information and will configure the protocol stack on vRP and UE-R side. In the case of left-hand side part of Figure 4-37, UE2 registers through UE1 (e.g. connects at L2 layer level to UE1) and then UE1 informs RTC/C3 (through RRC protocol or dedicated Relay Function) with the new topology. However, in the case of right-hand side part of Figure 4-37, UE2 registers directly to UE-R (e.g. connects at L2 layer level to UE-R) and then UE-R informs RTC/C3 (through RRC protocol or dedicated Relay Function) about the new topology.

The first 4 messages and configurations are identical for both left-hand and right-hand side of Figure 4-37, which show a UE1 registering to the NetWork (NW) through a UE-R and an RTC/C3 configuring vRP and UE-R with a single-hop configuration in order to control UE1 (as e.g. in Figure 4-34). This means that at this stage, UE1 connects directly to vRP and is normally under the control of vRP: the RRC UE1 is encapsulated in the PDCP UE1 and potentially transported through the PDCP of the UE-R to vRP (as e.g. in Figure 4-34). This corresponds to the "common part" of both (multi-hop and single-hop) situations represented in Figure 4-37.

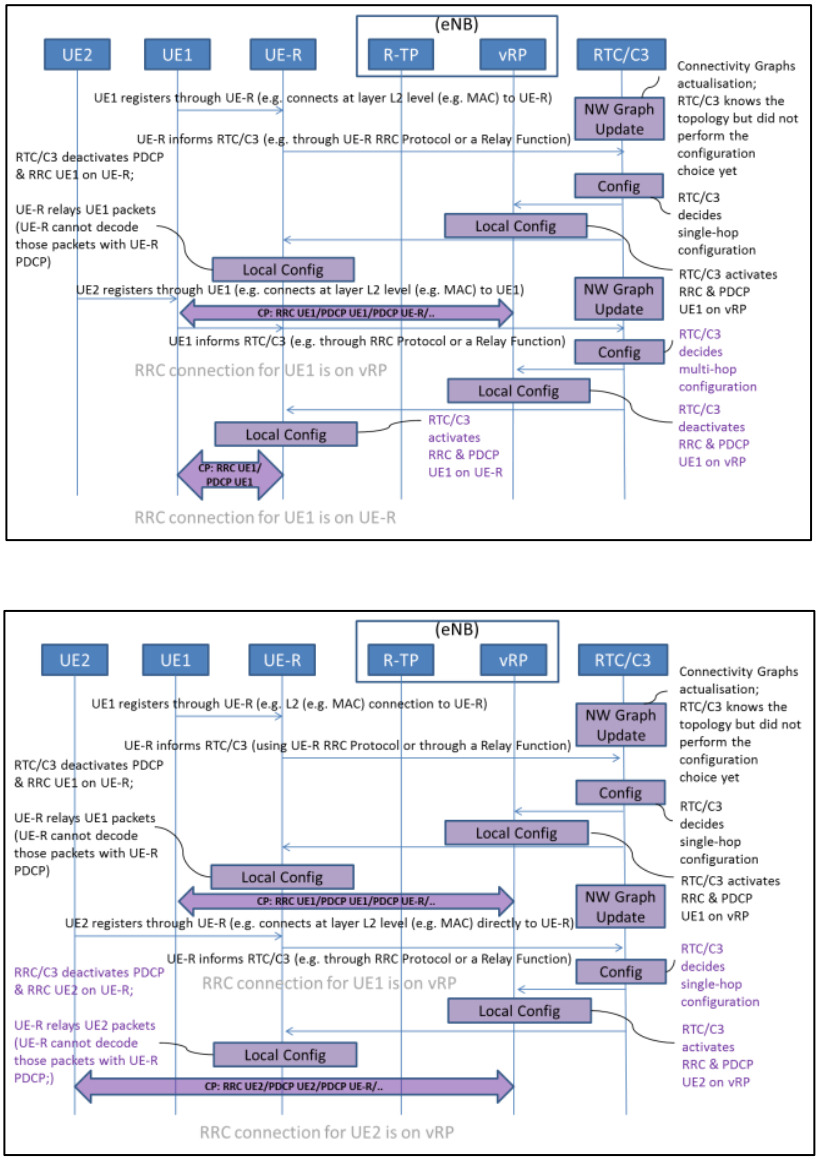


Figure 4-37: Multi-hop (up) & Single-hop (down) Configuration Examples

However, after this "common part", the difference between the examples from Figure 4-37 is further obvious:

- On the left-hand side of Figure 4-37 the RRC connection for UE1 is initially on vRP and then changes to UE-R as for multi-hop configuration represented in Figure 4-35 (while UE2 can be controlled by UE1 or UE-R), and
- On the right-hand side of Figure 4-37 the RRC connection for UE1 is on vRP and remains on vRP as for single-hop configuration represented for example in Figure 4-34 (while UE2 is controlled by vRP).

This reaction of RTC/C3 is motivated by the Network Graph update, which "sees" a new registered UE, a different network topology, and therefore decides to reconfigure differently the

system, providing a more relaxed (distributed) control or a centralized control (on vRP), depending the scenario or the real-time situation.

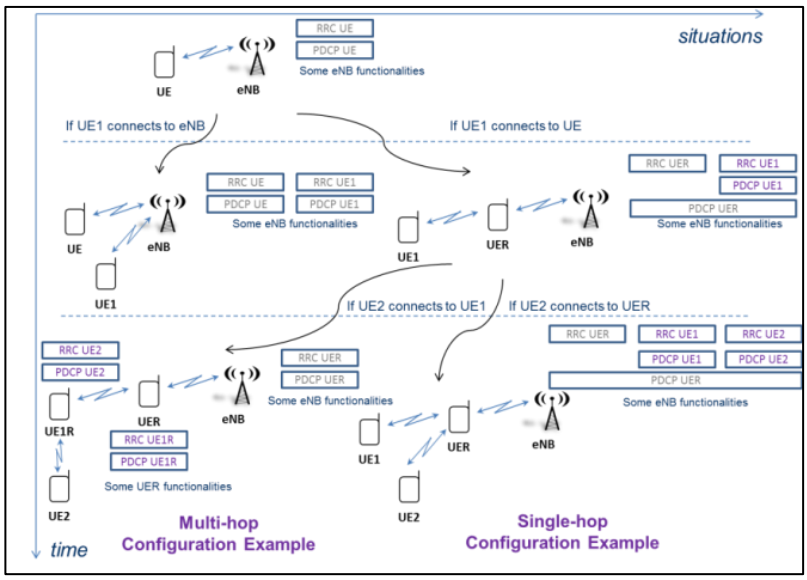


Figure 4-38: Multi-hop and Single-hop simplified descriptions

The above Message Sequence Charts (MSCs) can also be used to show the configuration complexity of different relaying schemes. For further exemplification, Figure 4-38 graphically shows how this situation could have happened.

While for multi-hop strategy, it has been decided to have distributed RRC and PDCP protocols to control UE2, UE1R and UER, for single-hop strategy it has been preferred to encapsulate RRC and PDCP of UE1 and UE2 directly in the PDCP layer of a UER, which means that the UE1 and UE2 are directly controlled by the vRP of an eNB. Moreover, the configuration may evolve with time as the RTC/C3 may decide to re-adapt to the real-time situation such as a (new) given network topology, if e.g. one of the UEs performs handover or another UE registers to the NW.

4.3.2.6 Evaluation and results

Evaluation methodology

A first evaluation can be performed through Message Sequence Charts (MSCs) as showed in previous subsection. This method is related to the implementation strategy and is therefore normally used to show the feasibility of different techniques. In this section, we present a second evaluation methodology: an inspection method based on latency evaluation of higher layer protocols. This study can be also extended to lower layer protocols.

Results

This paragraph describes why RTC/C3 has to continuously adapt the protocol stack and related required configuration in relation to topology changes, in order to master e.g. latency. The aim of this evaluation is to understand if the RRC protocol for remote UE control should be placed at the vRP side or rather at the UE-R side, depending on e.g. the number of hops in a multi-hop scenario (or other). We further focus on evaluation of control plane latency, in the case of extensions of the current LTE-A system. Currently LTE standard does not specify any mobility control for out-of-coverage users.

The main assumption used for the evaluation is that multiple instances of the protocol stacks are available at the same time over the same virtual machine on vRP and/or UE-R or any other equipment requiring fast and flexible reconfiguration. Another important assumption is the control plane latency of an L3 type protocol (RRC or NAS) for LTE is 10 ms for configuration messages; 10 ms for the measurement report; 10 ms for the decision & HO command to be send over an RRC protocol (see TS 36.211 [37], TS 36.331 [39]). Moreover, an extra 10 ms processing time is considered at UE-R level (for LTE radio frame decoding), which can be used to compute the overall end-to-end latency values between remote UE and eNB for different configurations.

In this deliverable we focus on the control plane latency (parameter configuration, processing time, reporting procedures and commands) as an upper bound for measuring the latency. The reason is that the latency induced by control plane may directly have an impact on the user plane, and can generate for example an interruption of the user plane data transmission during a time interval depending on the reconfiguration time required to re-adapt the network to environment changes.

Overall end-to-end latency should be minimized: it is known for instance that, for voice services, the latency must remain less than 100 ms (see TS 23.203 [11]), which means that multi-hop topology no longer meets the service prerequisites. It is therefore highly conceivable that the L3 signalisation over two hops between UE and vRP is no longer possible in order to transmit voice, because the control plane is no longer adapted to support user plane service requirements.

In M3.2 [5] we have further used Figure 4-39 to derive the end-to-end L3 latency values for best case and worst case scenarios from Table 4-18. In some situations, it is therefore better if the L3 control is not carried out by the eNB (vRP) and instead is directly done by UE-R, in order to correctly manage the mobility of an out-of-coverage user, but it depends on the UE-R degree of mobility. For example, in the case of one relay with one hop, if UE-R is moving (and not static as for previous case), the connection of UE1 with eNB may be lost for an undetermined duration.

However, as also seen in Figure 4-39, if UE-R is reconfigured by the eNB (through RRC protocol) prior to UE1 HO, some extra 30 ms latency are adding up to the total configuration latency, which can be explained by the fact that UE-R has to inform the remote UE about its new condition/situation (similar to a HO decision followed by a HO command with new access point measurement request, or new access point identity information, and new reconfiguration messages would apply) and then remote UE has to re-connect again to the (new) access point or network infrastructure.

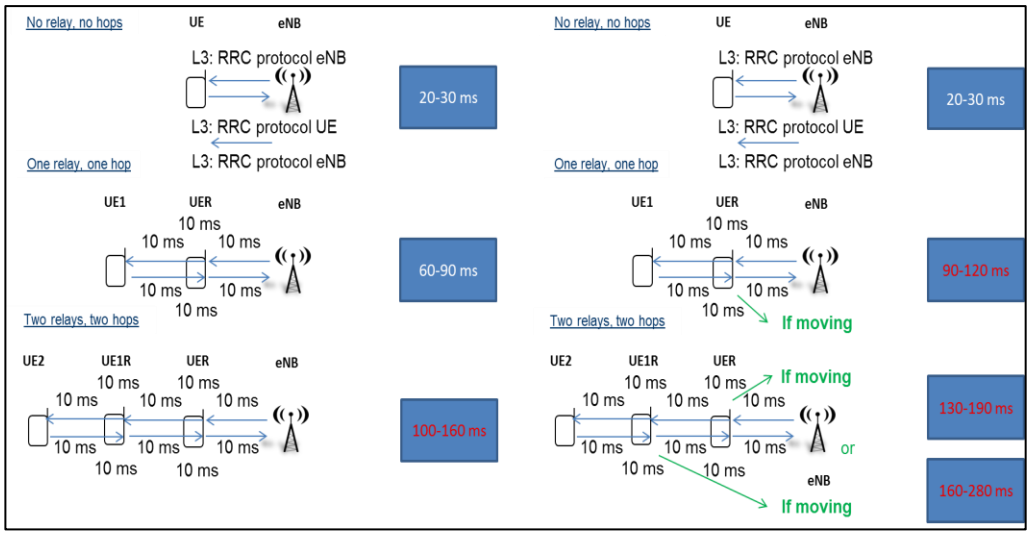


Figure 4-39: L3 latency evaluation: static and mobile (left-hand & right-hand side) scenarios

Table 4-18: L3 Latency Evaluation for Mobile Scenario

	Normal (legacy) topology without relaying or relays involved, without any hops	Single-hop topology, with one relay, and therefore one hop	Multi-hop relaying, with two relays, and therefore two hops
End-to-End L3 latency value (best case)	20-30 ms	60-90 ms	100-160 ms
End-to-End L3 latency value (worst case)	20-30 ms	90-120 ms (the relay is moving)	130-190 ms (the closest relay to the access point is moving) 160-280 ms (the farthest relay to the access point is moving)

Similar reasoning can be performed in the case with two relays and two hops, where if UER moves (but is reconfigured prior to UE2 HO) will increase the latency with 30 ms, while if UE1R moves (but is reconfigured prior to UE2 HO) will increase the latency with up to 90 ms, which also shows that the worst case is when the farthest relay from the access point is moving, which proves (again) that the control has to be distributed and in some cases located closer to the network edge.

The same logic can be applied for L2 level protocols, such as the PDCP sub-layer, because additional encryption can introduce additional delays. It is therefore necessary that RTC/C3, which has a global dynamic view of the entire network topology and parameters, makes a decision on the protocol stack configuration to be used (both L2, e.g. PDCP, and/or L3, e.g. RRC, NAS) and further on how to deploy the protocol stacks on the physical entities, e.g. directly on eNB, rather than on UE-R.

We must therefore make specific configurations to reduce the latency in the multi-hop case, to distribute the control between UE-R & eNB, to allow a faster re-configuration of the equipment. The RTC/C3 therefore has an important trade-off to deal with:

A. Implementing additional protocols on eNB (or rather on vRP in COHERENT flexible architecture) increases the complexity level at eNB side but it simplifies the complexity level at UE-R side (which is an advantage). Implementing additional protocols on eNB can also increase security for an out-of-coverage UE (another advantage).

B. However, implementing additional protocols on eNB side can increase overall latency (which is a disadvantage). In any case, it can also increase the interruption of service and the complexity of the HO procedure (for the detailed procedural steps of the HO procedure in this case, please see M3.2 [5]).

If we transmit all these configurations, measurement reports, commands and decisions between a UE and eNB through a UE-R, it will heavily increase system latency (and potentially also data flow). This is an additional reason for the fact that one might prefer to have a local RRC protocol at the UE-R level for the multi-hop procedure in order to control the out-of-coverage UE. The COHERENT architecture with R-TPs & vRP is a very good way to implement an eNB in a flexible and adaptable manner which will allow functionalities such as adding and erasing stack layers, protocols and parameters in the way previously proposed, which are transparent from the point of view of a user. Once activated, the RTC/C3 algorithm is executed sequentially within a loop, and has an overall adjustable loop execution speed which is sufficiently low to allow entities in the network to implement decisions made by the central decision-making body (RTC/C3). Typically, the execution time of a loop is of the order of hundreds of ms, while the configurations and the measurement reporting are of the order of tens of ms.

4.3.2.7 Conclusion

The role of this section was to show how the COHERENT architecture could be applicable to UE-to-Network relays, for single-hop and multi-hop communications. This flexible architecture allows not only to adapt the system to QoS requirements, but also to adapt the protocol stack accordingly and to allow a localized or a distributed control of the remote UEs. The main assumption used for this study is that multiple instances of the protocol stack of one single entity/equipment (vRP or UE-R) are available at the same time on the same virtual machine (on vRP and/or UE-R). This would allow for fast and flexible reconfiguration as already previously explained in WP2 and WP3, by taking into account in real-time the evolution of the system topology.

The system flexibility combined with the information from Network Graphs is allowing not only to adapt specific parameter optimization algorithms, but also to decide what protocols to use, in which situation and where to deploy them in the architecture. In this sense this contribution is complementary with other contributions in this deliverable.

Finally, we recall that this work is completely in line with 3GPP studies for REAR (Remote UE access via relay UE) which deals exactly with the idea of controlling Remote UEs, which was not possible so far with classical approaches provided by Rel-14 and previous ones.

4.3.3 D2D cooperation and relaying

4.3.3.1 Motivation and goals

Network-assisted Device-to-device (D2D) relaying is a promising way to provide continuous throughput performance for mobile users in cellular network [49][50][51]. The underlying idea is to exploit the availability of potential proximal D2D connections to setup beneficial multi-hop transmissions to/from base stations (BS). In [52], outage performance in a two-hop cellular network was analyzed. It was shown that outage gain provided by D2D relaying increases exponentially with the density of available relays. While standardization of D2D communication has been recently started, relay selection and interference management problems for D2D relaying have not been investigated properly in multi-cell networks. In the literature, D2D relaying has been addressed without considering the complicated interference between D2D relaying and cellular transmissions in multi-cell networks. As shown in Figure 4-40, when D2D relaying uses the same channel as cellular DL transmissions, a new type of interference generated by D2D relays emerges. It is caused to other D2D or cellular receivers in neighboring cells.

Compared to a multi-hop cellular network with fixed relays, a multi-hop cellular network with mobile D2D relaying is more difficult to manage. First, the active UEs and potential D2D relays are randomly distributed inside multiple cells. A BS controller needs to select a proper D2D relay for a UE who can benefit from 2-hop relaying, and to allocate proper resources to the D2D links. The following questions have to be addressed: *1) When to relay for a UE? 2) How to select a relay for a UE? 3) How to allocate resources for a D2D relaying transmission?* Second, in a multi-cell network where all cells adopt similar relaying strategies, D2D relaying decisions made in one cell would affect a neighbor cell's decision through changing the interference power distribution. A similar phenomenon is well known in the context of power control [53], but it has not been considered in the context of relaying. The neighbor cells would react to this change of interference by changing their relaying decisions. Accordingly, there is a feedback loop between the cells in the network, where decisions affect each other.

It is the objective of this section to study the interference interaction between cells in a network with D2D relaying, and to provide tools for a network controller to set parameters to achieve better cell-edge performance. We consider two-hop relaying for downlink communication. The interference characteristics for a multi-cell network with 2-hop D2D relaying depends on how much traffic is offloaded to two-hop relaying from one-hop DL transmissions, and the distribution of the selected D2D relays. Furthermore, the distribution of the selected relays depends on the

distributions of BSs and cell-edge UEs (which are identified by a network controller). We study the abstraction and characterization of the inter-cell interference in multi-cell networks with D2D relaying, based on a fluid model for interference.

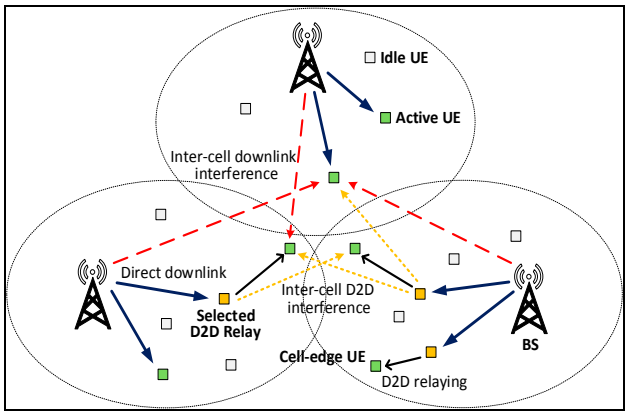


Figure 4-40 Inter-cell downlink interference from neighbour BSs and D2D interference caused by D2D relaying in a multi-cell network

4.3.3.2 Requirements

To manage D2D relaying in a multi-cell network, various parameters of the network need to be taken into account, for example, the distribution of BSs, the distribution of the UEs, the cell association principle, the relay selection principles and the resource allocation scheme for D2D relaying. The modelling and control of the inter-cell interference are difficult tasks for a real network setting. For example, the real distribution of BSs is often modeled as a stochastic geometric point process with a minimum inter-BS distance. Under such network, the downlink and D2D interference is very complicated. Network abstractions to capture the complicated interference interaction in the multi-cell networks with D2D relaying are essential for the D2D relaying management. To enable D2D relaying in cellular networks, D2D discovery, relay selection and resource allocation must be considered. We assume that UEs can perform channel measurements and report the channel information to the serving BS. A real-time controller located at the BS is responsible for relay selection and resource allocation for D2D relaying operations in its cell. We assume that a coordinated controller gathers information from the BSs on their interference measurements, and accordingly designs the control parameters to be used in the network. The BS controller has the responsibility to manage the interference in the network with D2D relaying. The ultimate requirement for the D2D relaying management is to provide satisfying data rates to cell-edge UEs while keeping the data rates for other UEs unchanged. To this end, we use a KPI called User Experience Consistency (UEC) as the network control objective function. UEC is defined as the ratio of the 5th percentile and mean throughput.

4.3.3.3 Abstracted parameters

We consider a system where D2D relaying happens in multiple cells using the same channel. D2D relaying transmissions in neighboring cells will affect each other if they use the same radio resources. The UEs close to the BS will use direct one-hop DL transmission, while two-hop D2D relaying may be used for network-to-UE transmission to cell-edge UEs. Without loss of generality, our analysis focuses on the cell with BS_0 located at the origin of the 2-D plane. We assume that all cells are similar. As shown in Figure 4-41, each cell is abstracted as a circular area with radius R_c. There is a ring of radius R_r ($0<R_r<R_c$), so that UEs at distance larger than R_r are served by D2D relaying. The D2D relays are assumed to reside in an annulus of inner radius r_1 and outer radius r_2. There is a homogeneous network of other cells around the cell at the origin. The transmit power of

all BSs is P_b, and the relaying UEs transmit with power P_d. We assume that there is a minimum distance D between the cell of interest and the neighbouring interfering cells. The BS-to-relay link uses a fraction β of the time and the D2D link uses a fraction $1-\beta$. This factor will affect the interference spillage to neighbouring cells.

We consider a UE_d located at distance d from the BS_0. When D2D relaying is not used in the network, the interference power for UE_d comes only from nearby interfering BSs, and we can use fluid model proposed in [54] to estimate the aggregate DL interference for UE_d. As in Campbell's theorem [55], the fluid model transforms an expectation of a random sum over the discrete point process (PP) to an integral involving the PP intensity in the distribution area. Assuming that the path loss exponent $\eta > 2$, then the aggregate DL interference is:

$$\mathrm{I_{dl}(d)} = \frac{2\pi\rho_{bs}P_bK_b}{\eta-2}(\mathrm{D}-\mathrm{d})^{2-\eta}, \qquad (4\text{-}7)$$

where K_b is the pathloss coefficient and η is the downlink pathloss exponent. As UE moves from cell centre to cell edge, the downlink interference from other BSs is increasing. The aggregate D2D interference is [56]:

$$\mathrm{I_{d2d}(d)} = \frac{M_d(D-d)^{2-\eta_d}(\phi(r_2)-\phi(r_1))}{(r_2^2-r_1^2)(\eta_d-2)}, \qquad (4\text{-}8)$$

where $\phi(r) = {}_2F_1\left(\frac{\eta_d-2}{2}, \frac{\eta_d+1}{2}; 2; \frac{r^2}{(D-d)^2}\right) r^2$, and ${}_2F_1(a,b;c;d)$ is the hypergeometric function, $M_d = 2\pi\rho_{bs}P_dK_d$ is a constant. Analysis of interference distributions from D2D transmission with arbitrary fading statistics can be achieved in closed form based on Poisson field analysis [89] . We assume that there is a probability pr that a D2D relay is transmitting in a given interfering cell. This D2D relaying activity factor is determined by the relaying strategy, such as R_r, β , r_1 and r_2. The interference experienced by UE_d from a neighboring cell (either from a BS or a D2D relay) is a random variable. The aggregate DL and D2D interference for UE_d is:

$$\mathrm{I(d)} = (1-\mathrm{p_r})\mathrm{I_{dl}(d)} + \mathrm{p_r}\mathrm{I_{d2d}(d)}, \qquad (4\text{-}9)$$

where p_r is the probability that D2D relaying is used in a cell for a specific resource block, we call it the D2D relaying probability and it depends on R_r, β , r_1 and r_2. The aggregate interference for UE_d is close to $(1-p_r)I_{dl}(d)$ if transmit power of BS is much larger than transmit power of D2D relay. In an interference-limited network, when D2D transmission probability p_r increases, the time-domain activity of BSs will decrease and the aggregate interference will decrease as the interference coming from BSs is on the average larger than interference from relays.

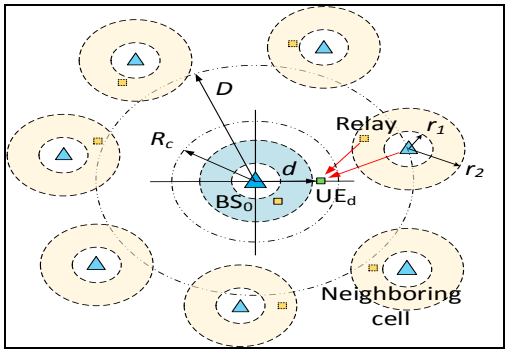


Figure 4-41 Network abstraction for D2D relaying

The relay selection and resource allocation for each cell-edge UE is based on throughput estimation using the measurements performed by UEs, while the D2D relay management (i.e. which UEs can be relay candidates) for all cell-edge UEs is based on the parameters shown in Table 4-19 Exposed parameters Table 4-19.

Table 4-19 Exposed parameters

Parameter	Description	Direction	Deployment
D	Minimum inter-BS distance	UL/DL	C3
R_c	Cell radius	UL/DL	C3
ρ_{bs}	BS distribution density	UL/DL	C3
$I_{dl}(d)$	The aggregate interference from all neighbor BSs for a UE located at a distance d from serving BS	DL	RTC
$I_{d2d}(d)$	The aggregate interference from all D2D relays in neighbor cells for a UE located at a distance d from serving BS	D2D	RTC
$I(d)$	The total interference for a UE located at a distance d from serving BS	DL/D2D	RTC

Table 4-20 Controlled parameters

Parameter	Description	Direction	Deployment
R_r	D2D relay distance, UE at a distance from its serving BS larger than which would use D2D relaying service	DL	C3
p_r	The probability that D2D relaying is used in a cell with a specific resource block	D2D	RTC

4.3.3.4 Generation and usage of Network graphs

The node in the network graph is a cell which is characterized by the population of active UEs, cell-edge UEs and relay candidates and their transmission powers. The edge in the network graph is the interference coupling happening between neighbouring cells, the interference consists of both DL interference and D2D interference. The network graph is depicted in Figure 4-42. There is a specific D2D relaying probability in each cell if the relay strategies described above section is adopted. D2D relaying decisions made in a cell will affect the neighbouring cell's decisions, and vice versa. The neighbouring cells affect each other in an iterative way in a feedback loop driven by changes in aggregate interference. In the equilibrium state of a homogeneous network, all cells should have the same D2D relaying probability p_r, and the same radiuses r_1 and r_2. With the optimal relay selection considered here, these numbers can be characterized in terms of a single radius R_r. A UE at distance d greater than R_r would use two-hop relaying. Assuming that the active UE and the D2D relay candidates are uniformly distributed in the cells and a cell edge UE at d with $d > R_r$ can always find a relay which is close to the optimal relay position for it, by assuming that there are large number of relay candidates to be selected. The distance of the optimal relay helping a UE at R_r would be r_1, and the distance of the relay helping a cell-edge UE at distance R_c would be r_2. The D2D activity factor p_r depends on the probability that the active UE is in the relaying region $d > R_r$, and on the ratio $1 - \beta$ of D2D transmission time. Based on these

abstractions, the network coordinator needs to determine an optimal D2D relay strategy which chooses a proper R_r for each cell in the network. If the network is homogeneous, the R_r is the same for all the cells.

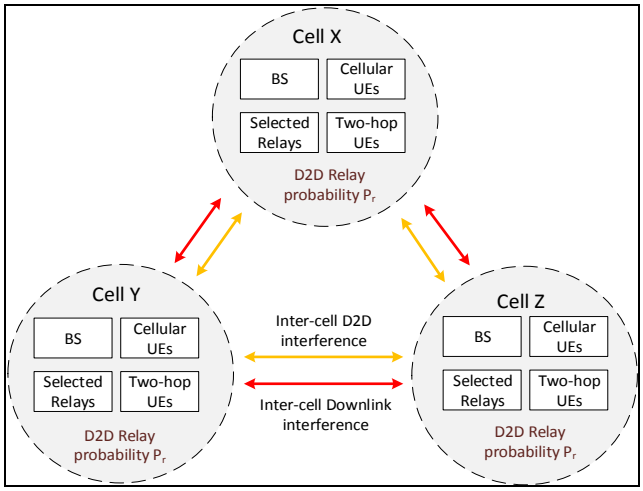


Figure 4-42 Network graph for multi-cell D2D relaying in downlink direction.

4.3.3.5 Algorithms in C3

To enable D2D relaying in cellular network, D2D discovery, relay selection and resource allocation must be considered. We assume that all the UEs can perform channel measurements and report the channel information to the serving BS. A real-time controller (RTC) located at the BS is responsible for the relay selection and resource allocation for D2D relaying operations in its cell. We assume that a coordinated controller (C3) gathers information from the BSs on their interference measurements, and accordingly designs the relaying distance to be used in the network. The BS controller has the responsibility to keep the probability of relaying at the target value. In this regard, C3 will optimize the R_r to be used by each BS, each BS then performs D2D relay selection for those UEs that can use D2D relaying service. We assume that the BS controller has the knowledge of D2D channel pathloss for close UE pairs, and the aggregate interference from other BSs. We also assume the same DL D2D relaying strategy is applied in all the cells, i.e. the same probability of D2D relaying, the same principle for relay selection (which determines the optimal relay for a cell-edge UE) and the same fairness metric for resource allocation (which is the objective function to be maximized for multiple UEs). The relaying decisions in one cell are made according to the current observed network state (e.g. aggregate interference at UE and distribution of relay candidates). The C3 is assumed to have all the parameters listed in

Table 4-19. It will determine an optimal relaying parameter R_r for the whole network. For the details of the algorithm, one can refer to D5.2 [7].

4.3.3.6 Results

In this section, we consider two multi-cell networks for DL D2D relaying, the triangular lattice (hexagonal cells). Simulation parameters can be found in D5.2 [7]. Each active UE located at distance larger than R_r from its serving BS would select the best relay from the relay candidates for its DL transmission. C3 starts search at $R_r = R_c$ with no D2D relaying allowed (i.e. $p_r = 0$). As R_r decreases, UEs at distance larger than R_r start to adopt two-hop D2D relaying. The average interference decreases as p_r increases.

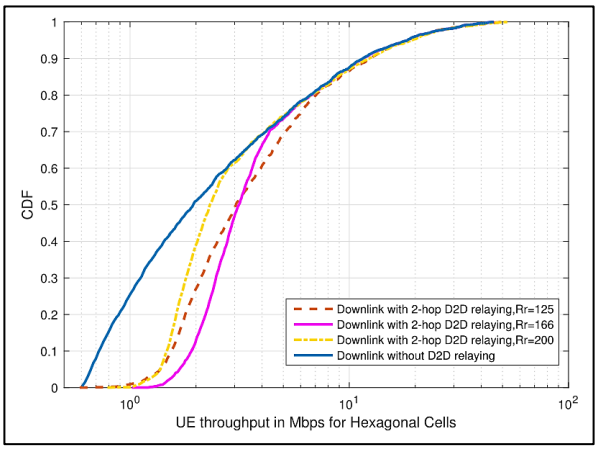


Figure 4-43 CDF of DL throughput

Figure 4-43 shows the Cumulative Distribution Functions (CDFs) of UE throughput performance for different D2D relaying strategies in the hexagonal network. A UE at a distance larger than R_r from its serving BS will adopt two-hop D2D relaying. D2D relaying helps to boost the performance of cell-edge UEs compared with the strategy that no D2D relaying is used. What's more, the setting of R_r = 166m yields the best cell-edge performance compared to the settings smaller or larger than it, and hence can achieve the best UEC performance. For R_r = 200m, fewer cell-edge UEs adopts two-hop relaying, which increases the aggregate interference as the relaying probability p_r is smaller compared to the strategy with R_r= 166m. For R_r= 125m, more UEs would use D2D relaying and the sum throughput for all active UEs is better than the strategy with R_r= 166m. However, a portion of cell-edge UEs will be affected by the increasing inter-cell D2D interference, which makes the cell-edge performance lower than the strategy with R_r= 166m.

4.3.3.7 Conclusion

We have considered two-hop D2D relaying to enhance end-to-end throughput in multi-cell downlink networks. This work was published in part in [56]. When managing Uplink D2D relaying, there is an additional complication arising from the power control of the users. Solutions for this are reported in [51]. For mmWave networks, relaying for coverage extension becomes more involved due to inherent directivity of the channel, while the gain potential also increases due to the high blocking probability and small coverage. A low-complexity relay & beam discovery and selection protocol for mmWave relaying is reported in [89]. In this section, we first study the characteristics of the aggregate co-channel interference when D2D relaying is applied in the networks. To capture the complicated interference interaction in the multi-cell networks with D2D relaying, we have proposed an analytical model with several parameters, including a minimum relaying distance and a relaying probability. These parameters are optimized for network-level management. Simulation results are provided to understand the interference characteristic and throughput performance when D2D relaying is applied. The analytical optimization of network parameters, based on this model, works as a good tool for network-level D2D relaying management.

4.3.4 Mobility study for high speed platform

As already explained in D3.1 [4], one of the use cases considered for high speed platforms usage and currently studied by 3GPP for 5G is Direct Air to Ground Communication (also known as DA2GC) which considers connectivity between airplanes and ground base stations (eNBs or gNBs). As we will show later in this section, the LTE communication system has to be properly reconfigured in order to become flexible enough with respect to new usage constraints such as high-speed platforms and air-to-ground propagation scenarios. For example, new measurements

strategies might be required with e.g. different measurement times and measurement periodicities as a function of mobility level. As we will see, not only handover might be affected but also system robustness (in general) is sensitive to frequency and timing offsets. High mobility and the system impact have to be properly evaluated in this new context. Moreover, through network graphs information and RTC for example (COHERENT controller placed near the access side), the system will permanently adapt, being therefore able to continuously provide robust services in both high mobility and low mobility scenarios.

In Figure 4-44 we have represented the block diagram of target system model, with C3/RTC entities and relayed users in airplane. The relayed users UE1 and UE2 (remote UEs) could benefit of services such as internet and/or voice, and could further access the terrestrial network infrastructure. The role of C3 would be to acquire useful information and to feed the network graph database, while the role of the UE-RTC and RTC would be to extract new information, and to configure new measurements.

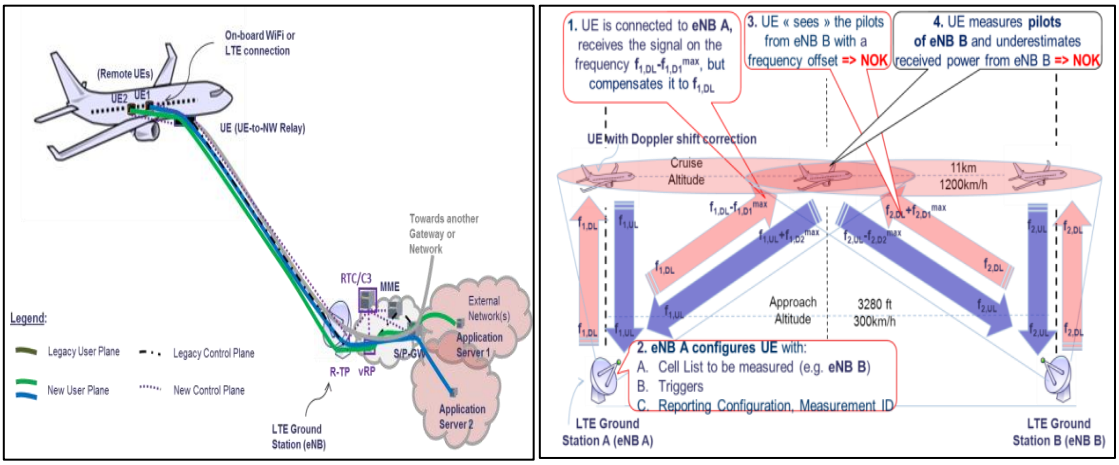


Figure 4-44 Block diagram of system & LTE measurement issues related to High Speed

4.3.4.1 Motivation and goals

As already explained in D3.1 [4], in the DA2GC scenario we can consider User Equipment (UE) mounted on an airplane and connected to an eNB (a similar scenario could be considered with an eNB mounted on an airplane, instead of a UE). This UE has the role of UE-Relay or UE-to-Network relay for the users inside the airplane requiring terrestrial network access. The UE with the role of UE-Relay and mounted on the airplane receives the signal from eNB A with an initial frequency offset (see step 1 from Figure 4-44). At this step, the UE is able to compensate at PHY layer the initial frequency offset, during the cell search process. The LTE system has been actually designed to support up to 500 km/h of mobility at 2 GHz range, but the combination of high speed and/or high mobility (extreme conditions) has been poorly studied by literature and the effects on system performance are not very well known. The serving cell eNB A may configure UE on airplane with the cell list to be measured (e.g. eNB B), with the proper triggers, reporting configuration and measurement ID (as seen in step 2). However, as represented in step 3, the UE "sees" the pilots from eNB B with another carrier frequency, resulted from both local frequency offset and relative Doppler shift ($f_{1,D1}^{max}$+$f_{2,D1}^{max}$) that is not able to compensate. For this reason, at step 4, UE measures eNB B and underestimates received power from eNB. This will result in incorrect and unreliable measurement values reported to eNB A, which will affect the HO. Not only that the reported measurement of eNB B will be incorrect, but it will also be reported to eNB A too late, and therefore a higher interruption time is expected. This interruption should be

normally avoided with the help of the new architecture provided by COHERENT, thanks to C3 and/or to a more robust waveform.

Through network graphs information and new COHERENT controller, the system will permanently adapt providing robust services for a very high mobility.

4.3.4.2 LTE configuration generalities

As presented in TS 36.211 [37], the Cell-specific Reference Signals (CRS) are pilot signals used by the eNB and transmitted in downlink. CRS are mapped on several resource elements, and the mapping is different according to eNB cell ID, the antenna number and time-frequency position of the resource element where the CRS are being sent. If two antennas are used, the eNB is not allowed to send any data on the resource elements used to transmit by any of the antennas. For other information concerning the bandwidth configuration, sampling frequency and other 3GPP information, please refer to the technical report "Mobility study for high speed platform" [40].

4.3.4.3 Requirements

With respect to the use case UC6.AG from COHERENT D2.1 [2], some of the challenges have been provided as follows:

- Provide standard unicast and multicast services to customers in airplanes in all phases of the flight with enhanced quality of experience.
- Support communications up to 1200 km/s.
- Support air-to-ground communications, i.e. aeronautical propagation channels.

And then, a complete set of COHERENT requirements have been further mentioned in previous deliverables: [R#2], [R#15], [R#31], [R#52], [R#58], [R#62].

Concerning the requirements, in high speed scenarios (e.g. for high user mobility) there are two noticeable issues: 1) the estimated power (or SNR) of the measured signal decreases with respect to Doppler shift, and 2) longer measurement durations combined with Doppler shift may severely deteriorate the estimation when signal correlation is involved at the receiver side for power estimation. In practice, the goal of COHERENT solution is to satisfy an imposed measurement accuracy requirement (e.g. +/-3 dBs or more).

4.3.4.4 Generation and usage of Network graphs

Obviously the impact of Doppler shift is depending on the configured system parameters but also on the employed method at the receiver side. The impact of the receiver and its parameterization are very important as they will affect the capacity of the system to react to different (extreme) channel conditions such as high Doppler, noise, interference or other. We show next that RTC/C3 is able to mitigate the impact of the Doppler shift by dynamically configuring the system with new parameters. For power measurements we use a correlation-based method in time-domain or frequency-domain. The time-domain power estimator we use over the signal received from neighbour eNB is further described in a generalized equation below:

$$\hat{P} = \max_{0\le\tau\le N_{FFT}+N_{CP}-1}\left(abs\left(\frac{\sum_{j=0}^{N}\left(\sqrt{P}\cdot r(j+\tau)\cdot s_{RS}^{*}(j)\right)}{N}\right)\cdot\frac{P_r}{P_{s_{RS}}}\right)^2 = \max_{0\le\tau\le N_{FFT}+N_{CP}-1}\left(abs\left(\frac{\sum_{j=0}^{N}\left(\sqrt{P}\cdot r(j+\tau)\cdot s_{RS}^{*}(j)\right)}{N}\right)\cdot\frac{1}{P_{s_{RS}}}\right)^2 \quad (4\text{-}10)$$

where N is the number of received samples (in time-domain), r is the normalized received signal (for simplification purposes we have considered that r is a signal with power P_r=1) and P is the received power of the target (e.g. neighbour) cell to be measured. s_{RS} is the reference signal pattern used for correlation and P_r/P_{sRS} is the total power ratio between the signal r (which contains both data and CRS pilots) and CRS pilot s_{RS}. The latter normalization is required to obtain the estimate of the total signal power (data plus CRS pilots) and not only the received power of CRS pilots.

Different variants of the time-domain correlation method exist, depending on the number of antennas used at transmitter side (neighbour eNB to be measured) and at receiver side (UE performing measurements). There are at least 3 possible strategies, but in this section we will present results only for two situations: 1) CRS information from only one antenna is used at the receiver side and 2) CRS information from two antennas is used at the receiver side (see Figure 4-45).

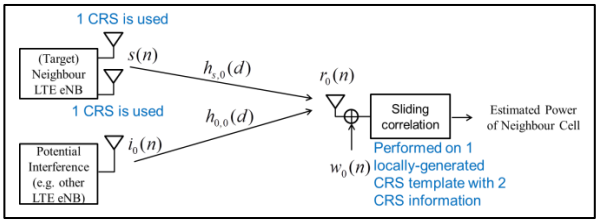


Figure 4-45 Simplified Model for Power Measurement using CRS – Tx with two antennas and Rx with one antenna

Therefore, if the neighbour eNB uses 2 transmit antennas, the test would be based on the locally-generated CRS template with two CRS information (the pilot reference signal pattern is for 2 transmission antenna ports):

$$T = \max_{0 \le \tau \le N_{FFT}+N_{CP}-1} \left| \sum_{n} r_0(n+\tau) s^*_{Pilot-2RS}(n) \right| \qquad (4\text{-}11)$$

The received signal is affected by noise, potential interference and Doppler, and therefore the estimated power will slightly change from the real received signal power. In this section we want to represent the effect of Doppler (and other parameters) on the measurement and on measurement method, and we will show that different configurations are required and several parameters have to be dynamically controlled by RTC/C3 as in the COHERENT framework, as a function of the environment changes (for example such as change of speed, change of frequency, etc.). The main statistics used for simulations results and further exploited by the network graphs are: measurement error; mean of measurement error; Root Mean Square Error (RMSE) of measurement error; Standard Deviation (STD) of measurement error; Probability Distribution Function (PDF) of measurement error. The measurement error expression is given by: $e_{\mathrm{dB}}[dB] = 10\log\hat{P} - 10\log P$, where P is the exact received power and $\hat{P}$ is the estimated power of the received signal affected by noise, interference, fading and Doppler. Measurement error is then computed over 10^3-10^4 simulation runs (i.e. realizations per each Doppler shift denoted as Nb_{sim}), and as a result of this computation, we can estimate the mean of measurement error μ_e, and other metrics and statistics such as RMSE or STD. For example, RMSE statistics are computed on the measurement error expressed in [dB] such as

$$\mathrm{RMSE} = \sqrt{\frac{1}{Nb_{sim}} \sum_{i=1}^{Nb_{sim}} e_{dB}(i)^2} \quad (4\text{-}12)$$

and STD of measurement error (i.e. standard deviation of the measurement error) is computed using the "population" Root Mean Square (RMS) equation

$$\sigma = \sqrt{\frac{1}{Nb_{sim}} \sum_{i=1}^{Nb_{sim}} \left(e_{dB}(i) - \mu_e\right)^2} \quad (4\text{-}13)$$

where μ_e is the mean of the measurement error [dB].

Other statistics can be also employed such as PDF of measurement error, PDF of reference Gaussian distribution having the same variance and mean as the measurement error (in order to show how Gaussian the measurement of error really is), or even more complex statistics such as Kurtosis, which is the fourth central moment of a variable (in this case the measurement error), divided by fourth power of its standard deviation minus 3. Statistics are used (for example) to represent the SNR estimation as a function of Doppler shift, and to allow the controller to know how the measurement estimation will deteriorate for a certain scenario, and how precise the estimation is. This information would allow for example the controller to adapt the system with new measurements configurations, and update the measurement parameters for example each time when the environment changes in terms of relative Doppler shift with more than 100 or 200 Hz (see offset constraint from TS 36.104 for eNB radio transmission and reception, and TS 36.101 for UE radio transmission and reception).

4.3.4.5 Abstracted parameters

Based on these previous evaluation assumptions, we can further imagine a system where the measurement configurations are performed by RTC, while the role of C3 would be to feed and provide the network graph database with more complete information and to compute more complex statistics, as represented in the table below.

Table 4-21 Exposed parameters

Parameter	Description	Direction	Deployment
$RSRP_b$, $RSRP_u$	Reference signal received power of BS/UE	UL/DL	RTC, UER-RTC
SINRb, SINRu	Signal to interference plus noise ratio of BS/UE	UL/DL	C3
Bb	Total PRB number of BS	UL/DL	RTC
Bu	Maximum PRB number of UE	UL/DL	RTC, UER-RTC
Relative Speed	Relative speed (or Maximum relative speed) of vehicle w.r.t. infrastructure	UL/DL	C3
Cyclic Prefix type	Normal or Extended Cyclic Prefix of the (OFDM) symbol, or other	UL/DL	RTC, UER-RTC
Pilot Power Boost value (if any)	The power of the pilots transmitted by eNB (if higher power than data channels)	UL/DL	RTC, UER-RTC
# of Tx antennas exploited for measurement	# of Tx antennas active at eNB side, 1 or more can be exploited for measurement (at UE side).	UL/DL	RTC, UER-RTC

# of Rx antennas used for measurement	# of Rx antennas active at UE side, 1 or more can be exploited. The measurement can be e.g. a mean, a max or other operation between 1 or more antennas.	UL/DL	RTC, UER-RTC
Type of Correlation Method	Time-domain or Frequency-domain correlation	UL/DL	RTC, UER-RTC
Carrier Frequency	The central frequency of the transmitter (e.g. eNB). The Bandwidth (centred on this frequency band) is derived from Bb parameter previously described.	UL/DL	RTC, UER-RTC, C3
Measurement Error	The difference between the estimated power and real power (evaluated or assumed).	UL/DL	C3
Mean Measurement error, RMSE and/or STD, PDF, CDF of measurement error	Statistical values obtained, estimated or measured over more than one measurement.	UL/DL	C3

Since in practice it is difficult to have a real-time estimate of the measurement error, this information could be for example used and exploited only at C3 level (from a database with predetermined simulations).

4.3.4.6 Results

In the table below, one may find a synthesis of different simulation parameters.

Table 4-22 System parameters used for simulation

Fixed/Variable Parameters used by Simulation	**Values used in the simulations**
BW	5 MHz
# of RB corresponding to BW	25
Carrier Frequency	2 GHz or more
Speed	Up to 1000 km/h or more
Tx # of Antennas	1 or 2
Rx # of Antennas	1 or 2
CP Type	Normal (7 symbols in a slot, CP1 for the 1st symbol and CP2 for the others symbols)
# of RE per RB	7x12
# of Pilot REs per Antenna (# of CRS REs per Antenna)	4
Doppler Shift	0 to 5 KHz (1/3 of subcarrier offset) or more
SNR LTE (signal to noise ratio of the measured cell)	0 to 30 dB or more
Noise Figure	11 dB
Channel Type	1-tap; 2-taps; 3-taps
# of LTE Transmitters	1 eNB
SIR LTE (signal to interference ratio of the measured cell)	10 dB, 0 dB, etc
Pilot to Data Resource Element Power Ratio	1 (0 dB)
# of Realizations per each Doppler Shift (to compute RMSE, STD, PDF)	1k-10k
(Power) Measurement Duration	1 symbol, 7 symbols, or more
Measurement Method	Time-Domain correlation or Frequency-Domain power estimation
CRS Pilot Modulation	QPSK
Data Modulation	QPSK
Sequence for Cell-Specific RS	For simplicity, Random Codes (we estimate this will not change the performance)

For the channel model we use single-path channel for most of the simulations, but the simulator can be extended to the multi-path fading channel model provided in 3GPP TS 36.521-1 V9.1.0, and 3GPP TS 36.141.

Potential Impact of COHERENT controller

Among the results we could cite for example the comparison of measurement estimation for 0.5 ms at 2 GHz carrier frequency for a speed up to 500 km/h and 1000 km/h respectively. Note that due to (potential) compensation with respect to the serving cell, the resulted Doppler shift can be actually up to 2 times higher with respect to the neighbour (target) cell, as explained in D3.1 [4]. The 0.5 ms measurement duration actually works very well for a maximum Doppler shift of 1000 Hz (e.g. at approximately 500 km/h, for 2 GHz; or e.g. at approximately 1000 km/h, for 1 GHz), but as the combination carrier frequency and speed increases, this is no longer the case. The role of the controller would therefore be to adapt the system with respect to the new conditions, when those conditions change. If we keep the measurement duration constant and increase the Doppler shift, the measurement error might considerably increase, which is reflected into the mean of measurement error which will differ from 0 dB value (0 dB measurement error corresponds to a perfect estimation) and will then reflect into the mean of estimated SNR.

Moreover, if the mean of the measurement error is 0 dB it does not mean that measurement is perfect. For this reason, in order to qualify the measurement over a larger number of realizations, it might be possible to use metrics previously mentioned such as STD and RMSE. The role of the controller would therefore be to adapt the system with respect to the new conditions (if measurement no longer precise and therefore uncertain). The reason for this important variation is related to the fact that during measurement duration the estimator based on correlation will capture more Doppler effect and therefore the phase of the received signal might change faster during the measurement duration. As a result of this, the estimation will not be reliable anymore. Another way to verify this statement is to choose a higher Doppler shift value and modify the measurement duration. One will therefore notice that at higher Doppler shift values is recommendable to use lower measurement durations while at lower Doppler shift values is recommendable to use higher measurement durations for better SNR estimate (especially for lower SNR conditions). In Figure 4-46 below we have represented the mean of measurement error for different measurement values. The lower impacted measurements at high Doppler are measurements performed on small segments (i.e. measurement duration is e.g. 1 symbol case).

The simulator allows also implementing a channel with fading. The fading will degrade even more the results, with 1-2 extra dBs, for both 1RS (the pilot signal reference pattern from 1 antenna port is used, i.e. 1 CRS information), and 2RS measurement strategies (the pilot signal reference pattern from 2 antenna ports is used, i.e. 2 CRS information).

Another comparison can be performed in terms of PDF, variance and mean. With dynamic provision of the measurement duration using COHERENT controller we can improve the measurement result. If we compare Figure 4-47 a) with Figure 4-47 c) for SNR=10dB with 0.5 ms (7 symbols) measurement duration and 0.5/7 measurement duration (1 symbol) respectively, we can see an improvement with more than 11dB in terms of mean of measurement error, and a smaller variance of error. Similar result can be found for higher SNR if we compare Figure 4-47 b) with Figure 4-47 d) for SNR=30dB. Higher the SNR is, better the measurement results are.

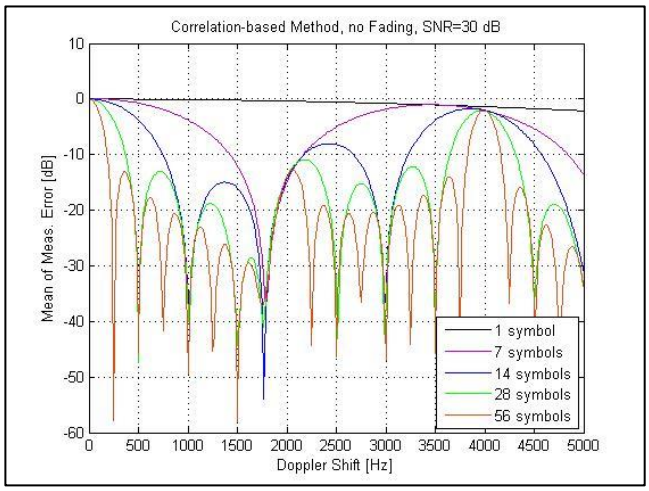


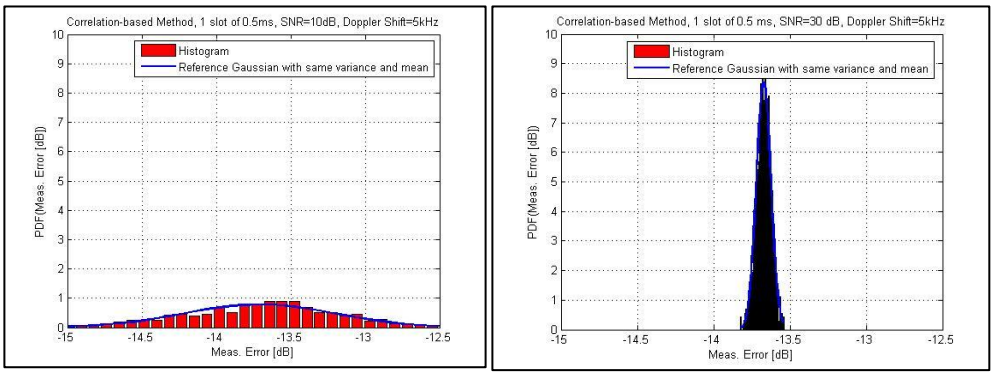
Figure 4-46 Mean of Measurement Error [dB] as a function of Doppler shift, for different measurement duration: 1 symbol (0.5 ms/7), 7 symbols (0.5 ms), 14 symbols (1 ms), 28 symbols (2 ms), 56 symbols (4 ms), for SNR=30 dB, 5 MHz LTE system with Normal CP

However, at lower Doppler shifts, and with some level of interference present, it is more interesting to use longer measurement durations. Longer the measurement is, lower the mean of measurement error is, with lower variance of measurement error, which finally translates in a better SNR estimation. For other information, please refer to the technical report "Mobility study for high speed platform" [40].

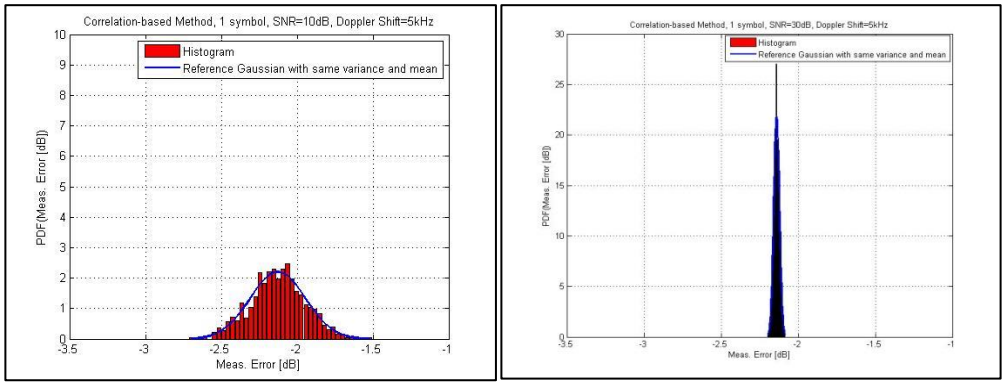
Even if the Doppler frequency is constant, the time is not and therefore the phase of the received signal will continuously change within the measurement duration that is used for correlation process. For this reason, if the power estimate is based on a correlation method between the signal affected by Doppler and a locally generated reference signal pattern (i.e. locally generated signal template), the final power estimate will not be correct anymore if the measurement duration increases. It is therefore important that the measurement duration is sufficiently small to diminish the Doppler impact, but at the same time sufficiently high to improve the SNR estimate. In other words, there is a compromise in terms of number of segments and segment length used for estimation. Another option would be to consider smaller measurement durations, and then to sum-up all estimation results and perform a mean on all multiple (smaller) measurements.

4.3.4.7 Algorithms in RTC/C3

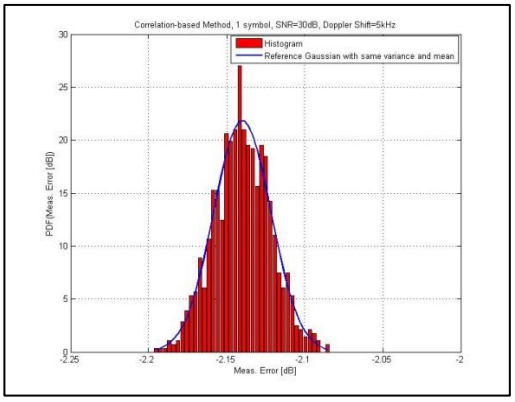
Based on all previous results, one can construct a strategy: for lower speeds and carrier frequencies use longer measurement duration (and more reliable in terms of SNR estimation), while for higher speeds and carrier frequencies use shorter measurement duration (and less reliable in terms of SNR estimation) but more frequent measurements (for example triggered by periodical measurements). The role of the RTC will therefore be to find the correct compromise between SNR estimation and estimation under (more or less) Doppler shift. We believe that this dynamic configuration will bring a lot of flexibility to the system and further allow different usages of the same system design, with different parameterizations. The goal of the RTC is to control the implementation of the correlation with a flexible measuring duration, and the condition exploited by the RTC is not to have an important Doppler shift contribution in the time frame used for e.g. power evaluation/estimation. The role of the controller is therefore to parameterize the PHY algorithm (choosing the length of the sequence), and/or to control the frequency of measurements.



a) b)



c) d)



e)

Figure 4-47 PDF for 5 KHz Doppler shift, without COHERENT controller a) 0.5 ms (1 slot) measurement duration, SNR=10 dB, b) 0.5 ms (1 slot) measurement duration, SNR=30 dB; with COHERENT controller c) 0.5/7 ms (1 symbol) measurement duration, SNR=10 dB, d) 0.5/7 ms (1 symbol) measurement duration, SNR=30 dB, e) zoom-in view of d)

Table 4-23 Controlled parameters

Parameter	**Description**	**Direction**	**Deployment**
Measuring Duration	Time required to perform one measurement, is defined as a maximum time constraint of a segment used for correlation process. It can be defined as a function of symbol length.	UL/DL/SL	RTC, UER-RTC
Measuring	Time required between two consecutive	UL/DL/SL	RTC, UER-

Period	measurements. It can be defined as a function of symbol length.		RTC
# of Measurements	The number of measurements required for an accurate estimate, for example precise power estimation.	UL/DL/SL	RTC, UER-RTC
Type of measurement	e.g. Time-domain measurement or Frequency-domain measurement, correlation or multiplication.	UL/DL/SL	RTC, UER-RTC

4.3.4.8 Conclusion

The role of the controller for the use case provided in this section is to find a reasonable configuration for SNR estimation under important Doppler shift. We believe that this dynamic configuration will bring a lot of flexibility to the system and further allow different usages of the same system, with different parameterizations. The COHERENT controller is essential for providing the correct measurement strategy, based on collected statistics and knowledge from the network graphs information. The main condition continuously checked by the controller is for example a limited Doppler shift contribution in the frame used for the power evaluation.

4.3.5 QoS Guaranteed Moving Relay in Self-Backhauled LTE

4.3.5.1 Motivation and goals

LTE is expected to be the next RAT for Public Safety (PS) communications and specifications items have emerged in this regard. PS networks have several specific requirements including reliability and resiliency which demand specific QoS to be addressed. In common PS scenarios, LTE eNBs may lose access to the EPC due to some outage or to constraints related to deployments. When that happens, they are not able to provide any service to their served users. This issue is addressed by 3GPP through the *Isolated E-UTRAN* concept that allows eNBs to continue providing minimal services for local PS users (TS 22.346, TR 23.797). Here, a further evolved concept is proposed of a new BS architecture for nomadic LTE networks that we called as *enhanced eNB* (e2NB). It embeds essential core network functions to ensure local services and has the ability to connect to other similar BSs leveraging virtual UEs and relay interface (i.e. Un interface) that is in-band with Uu interface to create a mesh network shown in Figure 4-48. The e2NB can support several new use cases where dynamic meshing among fix and/or moving BSs is required. For instance, it enables multi-hop in-band self-backhauling and create on-the-go wireless mesh network in a self-organized manner and it can be applied in the particular use case U3.MN "Coordination of rapidly deployable mesh networks" and U3.CE "Coverage extension and support of out-of-coverage communications" of D2.1 [2].

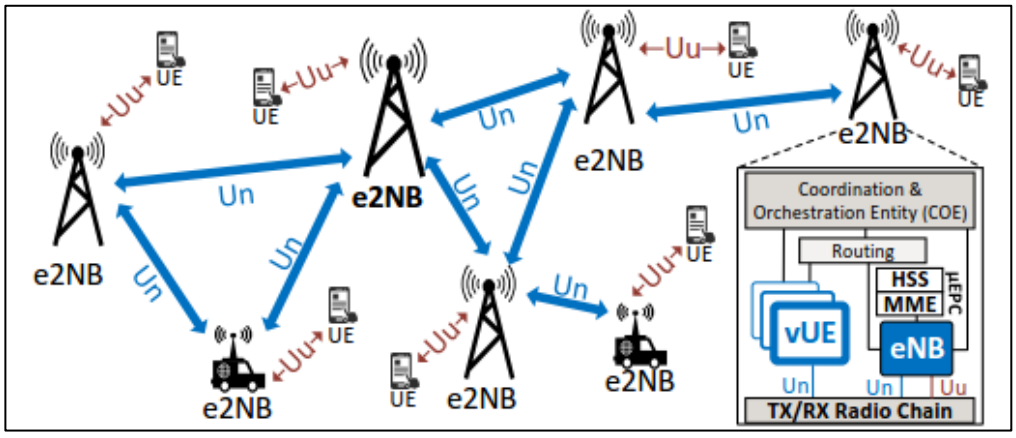


Figure 4-48 LTE mesh network based on e2NB with LTE backhaul

4.3.5.2 Requirements

Each e2NB can not only operate independently and host its own services, but also provide services to others through the self-backhauled mesh network using Un interface. To share the same radio source of the same carrier frequency between Un and Uu interface, the multicast-broadcast single frequency network (MBSFN) concept is utilized. In the resource allocation, sub-frames (SFs) are classified into MBSFN SFs for backhaul link transmission (i.e. e2NB↔e2NB) and non-MBSFN SFs for access link over Uu interface (i.e. e2NB↔UE). We provide an example of resource allocation in Figure 4-49 with 6 MBSFN SFs for Un interface (i.e. between eNB and vUE) and rest 4 non-MBSFN SFs for Uu interface (i.e., between eNB and UE). Note this example is based on LTE FDD mode with HARQ delay of 4SFs among DL reception and ACK transmission in the UL, so the SFs of DL and UL is offset by 4 SFs. Hence, we aim to allocate (a) MBSFN SFs for backhaul transportation, (b) next e2NB hop for backhaul relaying (c) relaying transportation direction (i.e. DL/UL), and (d) low-layer transportation resource (e.g., physical resource blocks, modulation and coding scheme) for both access and backhaul links.

Moreover, two different types of traffic are considered: real-time traffic with latency requirement (e.g. VoIP) and elastic traffic that is served in the best-effort manner. A timely-delivery of real-time traffic is essential to guarantee QoS for PS purpose. Hence, our approach has two objectives: (a) Guarantee the latency of real-time flows and (b) maximize the throughput of elastic flows. For the real-time traffic, we consider the VoIP traffic with maximum one-way-delay of 150 ms for 95% of the packet to ensure a quality call.

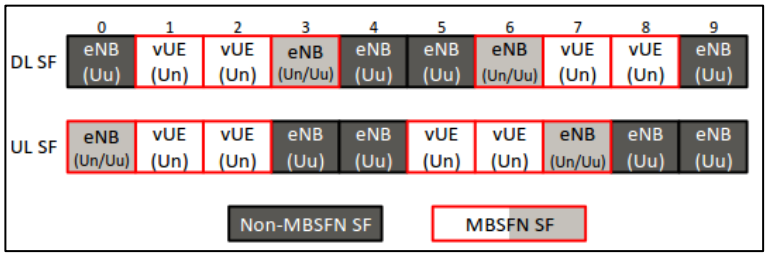


Figure 4-49 Example of MBSFN SFs use for in-band self-backhauling

4.3.5.3 Generation and usage of Network graphs

To form the network graph G, we firstly treat each e2NB as individual vertices (i.e. V_{e2NB}) and edges between adjacent nodes (i.e. E_{link}). Firstly, we apply the Dijkstra's algorithm to find the shortest path to route each traffic flow in terms of the number of hops, i.e. edge weights are 1. Since every node can reach its next hop through either UL or DL directions, we apply the expected waiting time of the packet in UL and DL queues as the edge weight to decide the direction (UL/DL) upon receiving incoming packets at each e2NB. After the routing and using the real-time traffic information, we can compute the link load on edge (u, v) as $load_{u,v}$ in terms of the number of real-time traffic bits to be transported in a SF for further node scheduling problem at the C3.

4.3.5.4 Abstracted parameters

Table 4-24 Exposed parameters

Parameter	Description	Direction	Deployment
SF_{rt}	Required SFs for real-time traffic of e2NB	-	C3
$SF_{e,D}$, $SF_{e,U}$	Required SFs for elastic traffic of e2NB (DL/UL)	-	C3
L_{SF}	Super-frame (SuF) duration	-	C3

G	Network graph G=(V_{e2NB},E_{link}) for scheduling	-	C3
$P_{x,w}$	Received signal power at e2NB w from x	-	C3
criteria(a, b)	Interference blocking criteria of adjacent nodes a, b	-	C3
SF_{MBSFN}	MBSFN SFs in a frame	-	C3
Q[x][p]	Queue size of (v)UE x with flow priority p	UL/DL	RTC
N_{PRB}	Number of available PRBs	UL/DL	RTC

Table 4-25 Controlled parameters

Parameter	Description	Direction	Deployment
SF_{TX}	SFs that e2NB shall transmit in a SuF duration	UL/DL	C3
PRB_{UE}, PRB_{vUE}	Allocated PRBs to UEs and vUEs	UL/DL	RTC
TBS_{UE}, TBS_{vUE}	Transport block size of UE/vUE in a SF	UL/DL	RTC

Role of RTC and C3 within the proposed hierarchical approach

Realizing an efficient mesh network on a single frequency band requires all nodes to share the resources of the same frequency band, and each transmitter becomes a potential interferer which limits the achievable rate. Further, there are more related issues that shall be considered in the meantime: (i) topology control, (ii) routing, (iii) link scheduling, (iv) interference measurement, and (v) power control. As all these issues are inter-dependent across different network layers; hence, a coordinated and cross-layer approach is provided to mesh e2NBs and guarantee per-flow QoS relying on the coordination between RTC and C3.

Ideally, all parameters are scheduled by the centralized C3 for both access and backhaul links. However, due to the limitation of real deployment and time-scale separation between C3 and RTC, the propagation of control messages over the backhaul links cannot be instantaneous. Thus, we can group parameters that shall be scheduled in a real-time manner (i.e. (c), (d) in 4.3.5.2 to be handled at RTC whereas others are allocated centrally in a larger time-scale benefiting from a whole network view (i.e. (a), (b) in 4.3.5.2 To enable such hierarchical approach, network information are abstracted by RTC to provide a simple but sufficient network information to the centralized C3. Detailed functionalities can be found in Figure 4-50 and the centralized C3 will do the flow routing, topology management, flow control and centralized node scheduling while the distributed RTC will do local control process and distributed link scheduler. Note the centralized C3 is beneficial when coordinating the interference between e2NBs whereas distributed RTC leverages the local link adaptation to allocate the radio resource in a real-time manner. Figure 4-50 also summarizes the required abstracted network information needed by C3 and the results provided by the C3 to each distributed RTC leveraging exchanged message between C3 and RTC.

Centralized node scheduler (CNS) and Distributed link scheduler (DLS)

The centralized node scheduler at C3 is executed periodically (i.e. the duration of a superframe) or on reaction to an event (for instance, a new traffic flow with real-time requirement) to coordinate e2NBs and allocate time resource for both Un and Uu interface, i.e. SF. During coordination among e2NBs, its goal is to allocate enough resource for each e2NB to fulfil the real-time traffic transportation. In contrast, the best effort traffic are taken into account only when the number of queued bits is significantly larger than the number of transported bits. Moreover, the interferer criteria is considered by C3 to not allow the transmission of some e2NB if they will generate interference to already activated e2NBs. Based on the allocated SFs, the distributed link scheduler at each RTC will execute at every SF and takes care of the frequency resource (i.e. PRB) scheduling among both Un and Uu interfaces. Our designed algorithm is to prioritize backhaul links (i.e. vUEs) over access links (i.e. UEs) as the former one can only reach 60% of peak rate (max number of MBSFN SF in a frame) compared to 100% for the legacy UE while also prioritizing real-time traffics over elastic ones to further reduce the unwanted latency. Further, a common round robin algorithm is applied for different UEs/vUEs.

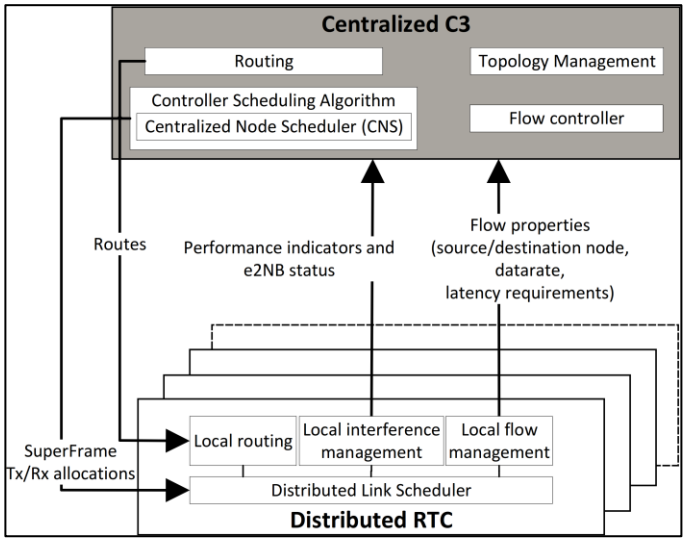
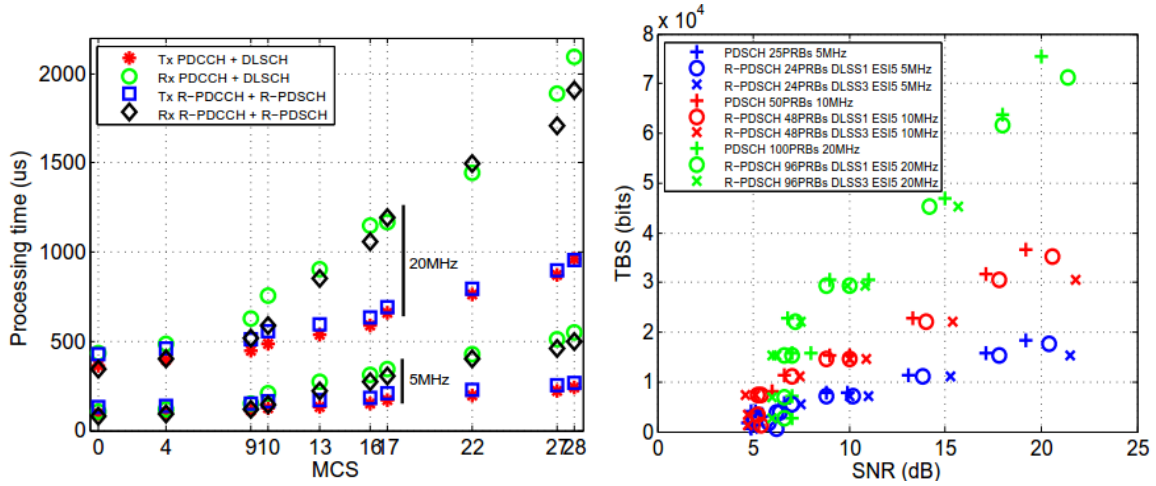


Figure 4-50 Relation and functionalities of RTC/C3

4.3.5.5 Results

Before providing the results of the applied C3/RTC algorithm, we firstly examine the impact of applying the physical channels over Un interface. In the Un interface, two new physical channels are introduced: R-PDCCH (Relay Physical Downlink Control CHannel) and R-PDSCH (Relay Physical Downlink Shared CHannel). The major change introduced by R-PDCCH is that it applies to the last 10-12 symbols within a SF whereas legacy PDCCH spans the first 1-4 symbol(s). Further, the R-PDSCH is similar to the PDSCH but uses less symbols and higher code rate. These properties induce that the procedures to decode R-PDCCH and R-PDSCH are different from the ones of PDCCH and PDSCH. Hence, we implement the R-PDCCH and R-PDSCH functionalities in OpenAirInterface and carry experiments to analyze the link-level performance in Figure 4-51. In Figure 4-51(a), the processing time taken by transmitter and receiver on both control and data channels among 2 system bandwidth (i.e. 5MHz and 20MHz) are shown and the physical channels of both Un and Uu interface take approximately the same execution time. In Figure 4-51(b), the link-level performance is examined in terms of the signal-to-noise ratio (SNR) level to successfully decode 75% of transported blocks. It can be inferred in figure that the efficiency of R-PDCCH/R-PDSCH is very close to that of PDCCH/PDSCH with slightly lower maximum achievable data rate. These results prove that the performance of the relay channels of Un interface is close to the one of Uu interface and promise a candidate for self-backhauling e2NBs.

After confirming the close performance of Un and Uu interface, we examine the performance of the proposed algorithm using C3 and RTC. In the evaluation, we generate an e2NB network topology with 7 e2NBs as shown Figure 4-52 (a), and each BS has 10 attached UEs initially. Our applied traffic patter comprises: (a) all UEs are randomly paired with each other UEs and establish a real-time VoIP call with a packet size of 20 bytes with an arrival rate of 20ms in both directions and (b) two fixed BS-to-BS elastic data transfers representing inter-sites (i.e. BSs) data transfers which happen in military and public safety application scenarios. In following, we make the comparisons between (i) the centralized approach proposed by Li and Ephremides [57] with corresponding modifications to fit in LTE system, (ii) the proposed algorithm but equally allocate MBSFN SFs to elastic and real-time traffic (denoted as the "Basic share algorithm"), and (iii) the full proposed algorithm (denoted as "proposed algorithm"). The approaches proposed by Li and Ephremides is to jointly allocate scheduling, power control and routing algorithm for TDMA wireless mesh networks.



(a) Processing time comparison (b) Minimum SNR to decode 75% of blocks

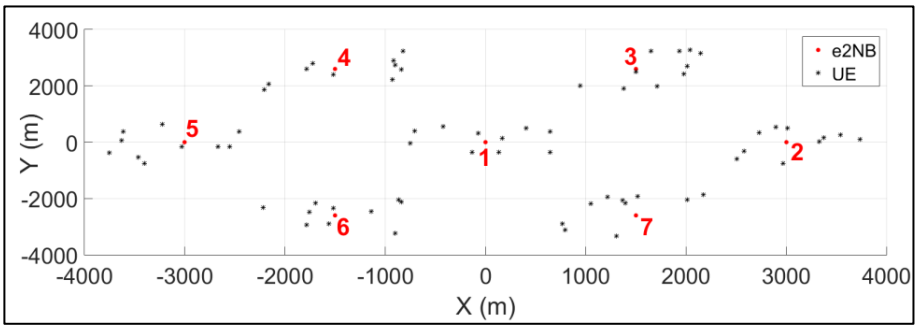
Figure 4-51 Results of R-PDCCH and R-PDSCH evaluation.

Then, the results of VoIP traffic flows are compared in Figure 4-52 (b) and both "proposed" and "basic" algorithms achieve 100% satisfaction to match 150ms end-to-end latency for the VoIP calls whereas the Li's approach fail to satisfy the latency requirement with 90% of VoIP flows cannot fulfill the requirement (i.e. 95% of packet shall with less than 150ms latency).

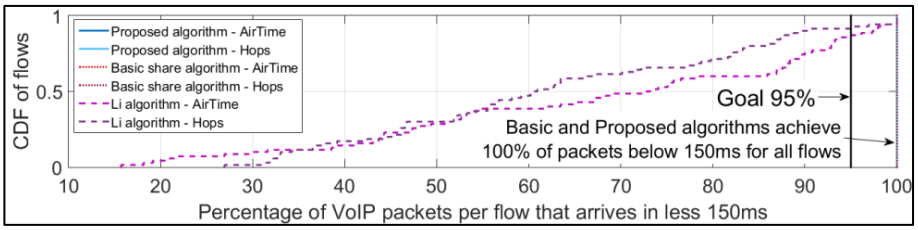
Such high dissatisfaction ratio is due to the lack of link adaptation capability and efficient multiplexing between e2NBs. Note we consider two packet sorting methods, denoted as number of hops (*Hops* in Figure 4-52 (b) which means to prioritize packets that with minimum hops toward its destination) or shortest deadline (*AirTime* in Figure 4-52 (b) which prioritizes packets with shortest deadline), do not make large differences to fulfill the deadline of the VoIP traffic. Then, in Figure 4-52 (c), we compare the cumulated data rate of elastic traffic flow among these three approaches and the two elastic flows are from e2NB#5 to e2NB#4 and from e2NB#2 to e2NB#4 (see Figure 4-52 (a)). The "proposed" algorithm, is the best as it can dynamically allocate SF between real-time and elastic traffic compared with the basic algorithm and Li's algorithm.

4.3.5.6 Conclusion

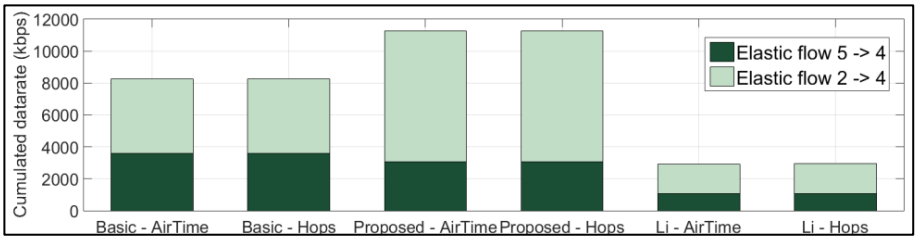
Our proposed hierarchical resource allocation approach can efficiently and practically realize an in-band LTE wireless mesh networks that leverages the LTE relay interface (Un) and takes into account QoS requirements of real-time flows. Such approach is based on the COHERENT architecture to separate the allocation algorithms into a centralized node scheduling (i.e. a global view of e2NBs over a large time-scale) and distributed link scheduler (i.e. local PRB scheduling in per-SF basis). Further, our findings are dual. Firstly, we show that in-band meshing of LTE base stations is possible without requiring intense modifications on current standard and physical layer design while keeping legacy UE support. Secondly, our approach is able to better fulfill QoS requirements of real-time flows while improving the global network throughput for elastic flows. These findings still exist in different topologies of network (e.g. line topology) or larger scale network (e.g. 19 e2NBs with 190 served UEs.).



(a) Hexagonal Topology with 7 e2NBs



(b) CDF of VoIP flows. (VoIP traffic and 2 inter-e2NB elastic traffic)



(c) Data rate of elastic traffic (VoIP traffic and 2 inter-e2NB elastic traffic)

Figure 4-52 Comparison of several approaches

4.4 Link performance

4.4.1 Building probabilistic network graphs with RAT-agnostic abstractions

4.4.1.1 Motivation and goals

Future 5G networks will experience highly dynamic networking scenarios created by node mobility, traffic heterogeneity and dynamic interface selection in dense heterogeneous networks. Such a scenario makes necessary the representation of network state which is comparable across RATs and can be used to obtain probabilistic guarantees over QoS instead of using just averages. Such representations are required to fulfil sophisticated coordination and control of resources and satisfy individual user equipment (UE) service demands by dynamically updating user associations.

To address this gap, we introduce a novel technology agnostic abstraction based on probabilistic modelling of *attainable throughput*. In D3.1 [4], we propose a WiFi specific link state representation, which we have now extended, by introducing attainable throughput, to a more general RAT-agnostic abstraction. Our general definition of attainable throughput offers a comparable way of predicting link performance across heterogeneous wireless infrastructures, which we exemplify in detail for WiFi and LTE radio access technologies (RAT). An empirical

CDF of the estimated instantaneous attainable throughput (over short time windows in reference to the rate of change of network state) serves as the basis of its probabilistic representation, allowing the controller to evaluate the probability with which a candidate link can satisfy the QoS requirements of the user. We shall use the general term served node to describe a UE or a client with any radio access interface, and the term serving node to describe an eNodeB or an AP.

4.4.1.2 Architecture and requirements

The aggregated network state is made available at the controller in the form of a network graph

[3]. The network information function (NIF) is a distributed function that aggregates input metrics and provides a RAT agnostic probabilistic abstraction. A local instance of NIF is instantiated on each serving node and the controller, which could be RTC or C3 [3]. The RTC and/or C3 use the network graph as input for control applications (ex. load balancing, QoS based traffic steering) running over heterogeneous networks.

In the current modelling approach, we assume that the data flows in the network are UDP and that each served node has one downlink data stream. The WiFi standard used in our work is 802.11g with 802.11k to communicate neighbour information to the associated AP.

4.4.1.3 Approach

Considering an application where the user experience at a served node is throughput driven. A certain level of instantaneous throughput is necessary to meet the requirements of the application. We perform passive observations in the network accumulating relevant metrics over measurement time windows, W_m. Measured and estimated throughput is evaluated as an empirical CDF over a sliding window, W_s of an appropriate number of measurement time windows. A threshold on this instantaneous throughput on the CDF provides a probability of meeting the throughput requirement. Non-parametric representations with the use of empirical CDFs are used to keep the solution general in a dynamic heterogeneous scenario.

Our work currently considers downlink modelling of attainable throughput with UDP traffic.

A controller dynamically monitors the QoS on links and updates user associations as required to maintain QoS. The choice of target serving node in the event of an association update would be obtained by selecting links whose attainable throughput estimations satisfy the probabilistic requirements for QoS.

4.4.1.4 Abstracted parameters

We present here the base metrics and the estimation model used for the RAT agnostic representation of attainable throughput in both WiFi and LTE RANs.

Metrics

We begin with the metrics aggregated for WiFi RAN.

RSSI: The signal to noise ratio (SNR) obtained from RSSI and noise at the receiver informs about the setting of sending rate or modulation and coding scheme (MCS) at the transmitter, based on target thresholds for bit error rates (BER). Thus, using SNR and BER thresholds, the MCS that will potentially be used on the link can be estimated[68].

Packet delivery ratio (pdr) is a measurement of packet reception rate. While RSSI captures the effect of fading on a link, it is known that it does not capture the effects of packet collisions from random multiple access [41]. When a well-functioning sending-rate-control algorithm at the physical layer, sets the MCS to obtain a target low BER, the losses from fading are low and the drops are dominated by collisions. Assuming this collision limited scenario, packet delivery rate

can be measured at the access point by counting data packet transmissions and their acknowledgements.

Arrival rate (arr): The arrival rate in packets per unit time of each flow is measured at a switch on the backhaul link.

Packet length (p): The average packet length of the packets in a flow is also measured at a switch on the backhaul link.

Number of active served nodes (N_a): The number of served nodes that have active data flows in the current measurement time window as measured at a switch on the backhaul link.

The metrics we use for the representation of LTE RAN are detailed below.

RSRP: Reference signal received power is a measure of received signal strength expressed in dBm. RSRP measurement along with noise at the receiver provides SNR (represented as CQI), which informs the setting of sending rate or MCS (m) at the transmitter using SNR thresholds for target transport block error ratio (TBLER)[69].

The arrival rate, packet length, and number of active served nodes are measured for LTE in the same way as it is for WiFi, at a switch in the backhaul of the network.

Estimation

In WiFi networks the multiple access scheme used is random access using CSMA/CA in both uplink and downlink, with the wireless spectrum being shared over time between uplink and downlink transmissions. As mentioned before, we model the downlink transmission of UDP traffic. In downlink, the AP MAC performs first-in-first-out over the packets in its input queue (round-robin over the packets). Since the packets are in a single queue, the fraction of packets in the queue that belong to a certain served node is proportional to the packet arrival rate of that node. We use the basic representation of saturated WiFi throughput in [90] and extend it by using arrival rate as a weight for each active user's flow. We also extend it to represent unsaturated network scenarios, and summarize the estimation equations below.

$$\lambda^{i}_{sat,k} = \left(pdr^{i}_{k}\right) \times \frac{arr^{i}_{k} \times p^{i}_{k}}{\sum_{i \in S_a} arr^{i}_{k} \times \left(\frac{p^{i}_{k}}{m^{i}_{k}} + DIFS + SIFS + T^{i}_{ack,k} + T_{phy}\right)}$$

$$\lambda^{i}_{unsat,k} = arr^{i}_{k} \times p^{i}_{k} \times pdr^{i}_{k}$$

$$\lambda^{i}_{k} = \min\left(\lambda^{i}_{sat,k}, \lambda^{i}_{unsat,k}\right) \qquad (4\text{-}14)$$

λ^{i}_{k} is the estimated attainable throughput in measurement time window k evaluated for served node i. S_a is the set of active served nodes in k. $T^{i}_{ack,k}$ is the time to send a standard length MAC acknowledgement using the appropriate m^{i}_{k} as specified by the data. T_{phy} is the time to send the physical layer header. DIFS and SIFS are standard specified constants.

LTE networks on the other hand have scheduled access to spectrum resources in both uplink and downlink. Downlink throughput achieved at the served node in an LTE network depends on the MAC scheduler at the eNodeB used to assign resources, and its fairness objectives. The proportional fair (PF) scheduler provides a metric of throughput fairness among the served nodes in the network [70]. In practice, most comparative experiments on LTE networks are performed

using PF since it is a widely used scheduler. While there are other schedulers, here, we employ PF scheduling in line with common practice [71].The throughput estimation model is specific to PF and can be extended for other schedulers based on their resource allocation objectives. We use prior work presented in [71] to estimate the share of resources allocated (in transmission time intervals (TTI)) to each served node flow (Y^i) based on the requested resource (X^i) and the current network condition. We extend this to estimate downlink throughput at the served node (λ^i). Equation (1) classifies served node flows as low-rate (LR) or high-rate (HR) based on the fraction of available resource a user is requesting. TBS is the transport block size sustained by the link and is a function of the eNodeB's bandwidth and m^i. Using equation (2) for the assigned resources in equation (3) provides the attainable throughput in a measurement time window for each active served node.Table 4-26 summarizes the final set of abstractions that are exposed and the control parameters that can be set using the exposed abstractions.

$$X_k^i \leq \frac{T_{frame}}{N_{a,k}} \rightarrow X_k^i \in LR \qquad X_k^i > \frac{T_{frame}}{N_{a,k}} \rightarrow X_k^i \in HR$$

$$Y_k^i = \begin{cases} X_k^i, if X_k^i \in LR \\ \frac{1}{|HR|} min\left(X_k^i, T_{frame} - \sum_{LR} X_k^i\right) if\ X_k^i \in HR \end{cases}$$

$$\lambda_k^i = min\left(Y_k^i \times TBS_k^i, Y_k^i \times p_k^i\right) \qquad (4\text{-}15)$$

Table 4-26 Exposed and control parameters

Exposed Parameter	**Description**	**Direction**	**Deployment**
$P(\lambda > \lambda_{th})$	Probability of attaining QoS specified throughput at served node on a candidate link.	DL	RTC/C3
$P(\rho > \rho_{th})$	Probability of achieved throughput on a serving link satisfying the QoS requirement.	DL	RTC/C3
$F_\lambda(x)$	Probability mass distribution of attainable throughput.	DL	RTC/C3
$F_\rho(x)$	Probability mass distribution of achieved throughput on associated link.	DL	RTC/C3
Control Parameter	**Description**	**Direction**	**Deployment**
BS<links>	List of acceptable links or interfaces based on the QoS requirements of a served node RLC or WifFi MAC layer.	DL	C3

4.4.1.5 Verification

The attainable throughput estimation model for WiFi and LTE have been evaluated using simulation networks implemented on NS3 [72]. The output of the estimation model is evaluated against the ground truth achieved throughput. The approach of using CDF of attainable throughput to do interface selection and make mobility management decisions has been evaluated in detail in our technical paper that has been submitted and is under peer-review [92]. We compare our approach against existing state-of-the-art and demonstrate that our approach performs better at maintaining probabilistic user performance guarantees. We show that our approach results in 20% lower performance violations without increased handover cost. Due to the reason of brevity, this

extended evaluation has been added as a reference [92]. We have also made the code for WiFi and LTE NIF available on GitHub [93].

Results for WiFi attainable throughput estimation presented in Figure 4-54-a shows absolute attainable throughput estimation error as a fraction of ground-truth versus increasing measurement window size. 25th, 50th, and 75th percentiles of the error have been plot. Estimation errors are below 10% for measurement window sizes larger than 500ms. Estimation error is low for smaller number of served nodes, when the network is more likely to be in an unsaturated state. Moreover, we see that with increasing the number up to 20 served nodes, the estimation error rate increases as an effect of the network getting closer to saturation. We also observe that the variation in the estimation error reduces as the measurement window size is increased. When the measurement window is small in terms of the number of observable packets, the estimation error will be higher due to a potentially insufficient number of observations in the chosen measurement time window. These observations indicate the measurement window size should be set large enough to estimate throughput with required accuracy.

LTE has scheduled resource access, wherein the PF scheduler, in our case, allocated resources to served nodes as required based on fairness metric. This is a point of contrast between LTE when compared to WiFi, which performs random access. Figure 4-54-a plots the same quantities as Figure 4-54-b, but for LTE attainable throughput estimation. Estimation error reduces with increasing measurement window size for the same reason as we stated previously. The differentiating observation here is that the estimation errors are lower than that of WiFi. The PF scheduler in LTE actively assigns resources to served nodes to maintain fairness in throughput and hence maintains a more stable throughput at each served node.

The empirical cumulative distribution function of attainable throughput is shown in Figure 4-56 for WiFi and LTE. The vertical line is the QoS required throughput threshold (λ_{th}) and the horizontal line is the probability $P(\lambda \leq \lambda_{th})$. It shows how the estimated performance varies as the network changes with increasing number of served nodes. Using such a probabilistic threshold on D_{λ}, the controller can evaluate whether or not a candidate link will meets the QoS requirements of the served node.

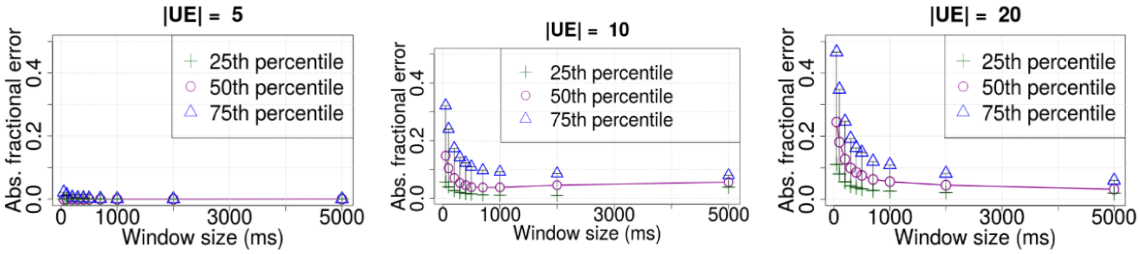


Figure 4-54-a: WiFi, attainable throughput estimation error as a function of measurement time window size

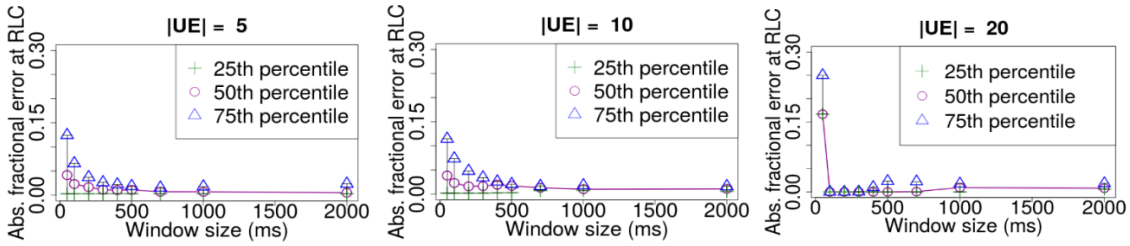


Figure 4-54-b: LTE, attainable throughput estimation error as a function of measurement time window size

4.4.1.6 Conclusion

An attainable throughput estimator for WiFi and LTE throughput estimation was evaluated. The approach enables automated controller decision making, based on probabilistic service guarantees for served node flows. For delay sensitive data flows such as VoIP, the network graph needs to be extended with a probabilistic estimate of delay, which we shall address in future work.

4.4.2 Distributed antenna systems for throughput improvement

In the recent years, DAS has gained interests due to its viable solutions for cell coverage extension, high frequency re-use and improved latency [74][75]. Multiuser DAS is one of the most promising techniques for 5G systems, especially for indoor [76] or outdoor hotspot coverage [77]. DAS mainly differs from a conventional collocated antenna system, where all the antennas are collocated at the base station, by employing geographical distribution of RRHs. The RRHs are connected to a centralized processing unit using dedicated cables, usually optical fibers. DAS has proven to offer viable solutions to the challenges such as throughput improvement, call blocking rate reduction, coverage improvement and reduction in transmit power [78][79].

4.4.2.1 Motivation and goals

Significant work [80][81][82][83] on DAS to understand the advantages and possible deployment strategies of DAS in comparison with collocated antenna systems has been done. The work proposes a set of exploitation methods and modifications of LTE PHY processing blocks and the relevant LTE procedures at the eNodeB for supporting the functionalities of DAS in a LTE system. Selection transmission (ST) [78], a simple DAS technique, which is based on selecting a single RRH out of all the RRHs for transmission has gained considerable interest in DAS, as it reduces the other cell interference (OCI) and retains the benefits of spatial diversity. In [85], a study on different schemes for selection transmission to increase the sum rate has been done in the case of a single cell multi-user downlink (DL) distributed antenna systems. However, in order to improve the sum rate of the cell, ST based pairing scheme, which pairs the UEs to the RRHs based on their distances or path losses is considered for our work. The ST based pairing offers opportunities to explore the re-use of time-frequency resources in the same cell. Probability of coverage metric is introduced to analyze different DAS scenarios. This metric provides reliable means to exploit the opportunities for pairing depending on the expected SINR at different UE positions.

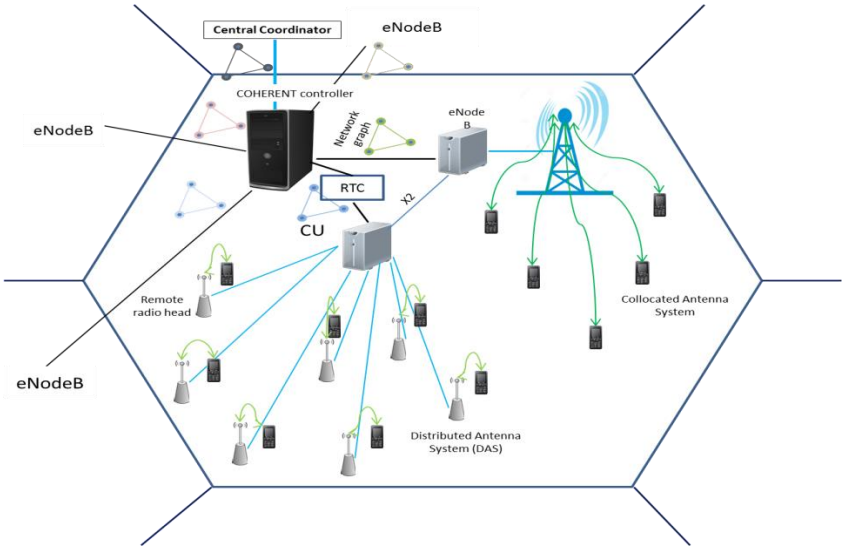


Figure 4-55 DAS EnodeB with Centralized Unit in HMN

Figure 4-55 shows a typical HMN scenario, i.e. DAS coexisting with a regular eNodeB, where the DAS eNodeB / central unit (CU) does all the processing of higher layers until physical layer and sends the time domain baseband signals to the RRHs via optical cables making use of baseband transmission standards like common public radio interface (CPRI) or open base station architecture initiative (OBSAI). Figure 4-55 also shows the general idea of having paired selection transmission technique where the RRH serves the nearest UEs. The simplest technique of ST based pairing as described in [85] is to pair UEs with RRHs based on the distances or path losses. In order to increase the sum rate using ST, a set of RRHs can use the same time, frequency and space resources thereby causing tolerable intra-cell interference. In some real-life scenarios, where the UEs are mostly stationary or less mobile and the need for sum rate is very high, for e.g. stadiums, theaters, shopping malls etc., one can pair the UEs with the RRHs so that the same scheduling resources of the pair can be reused in another pair quite far, i.e., where the signal to interference plus noise ratio (SINR) is high enough to decode the channels.

The targeted goals for COHERENT using DAS are per-user throughput improvement using high frequency re-use in different RRH distributions and analyzing the interference based on the probability of coverage metric. This work is published in [58].

4.4.2.2 System Model

We consider that the cell consists of ***N*** RRHs and ***K*** UEs distributed over the cell. Each RRH may have multiple antennas as defined in the LTE standard, but for the sake of simplicity the single antenna transmission is considered. The received signal at any i^{th} UE within the cell from j^{th} RRH can be expressed as

$$y_{ij} = h_{i,j}\, x_j + z_i \qquad (4\text{-}16)$$

where $h_{i,j}$ and x_j denote the coefficients of small scale fading and signal transmitted by j^{th} RRH respectively, z_i represents the noise at the receiver. The probability distribution function of the received power, $r_{i,j}$, based only on simple rayleigh fading channel at the i^{th} UE is given by:

$$f_r(r_{ij}) = \frac{1}{2\gamma^2} \exp\left(\frac{-r_{ij}}{2\gamma^2}\right) \qquad (4\text{-}17)$$

where $2\gamma^2 = \frac{P_j}{{d_{ij}}^{\alpha}}$ represents the mean received power of the signal, where we assume that the transmitted power falls off based on d_{ij}, i.e., the distance between the RRH port j and the i^{th} UE and path loss exponent α. $E\left(x_j x_j^*\right) = P_j$ represents the transmitted power of j^{th} RRH.

Probability of coverage metric

In this subsection, we introduce the probability of coverage metric [86] for analyzing different selection based pairing.

In a interference and noise limited system, i.e., when all the RRHs are transmitting, the received signal at the ith UE is

$$y_i = h_{i,i}\, x_i + \sum_{j=1, j\neq i}^{N} h_{i,j}\, x_j + z_i \qquad (4\text{-}18)$$

Probability of coverage based on SINR is given by

$$p_c = P\left(SINR_{i,j} > \Theta\right) = P\left(\frac{r_{ij}}{I+\eta} > \Theta\right) = P\left(r_{ij} > \Theta(I+\eta)\right)$$

$$= E_I\left(\int_{\Theta(I+\eta)}^{\infty} \frac{{d_{ij}}^{\alpha}}{P_i}\, exp\left(\frac{-r_{ii}\,{d_{ij}}^{\alpha}}{P_i}\right) dr_{ii}\right) \qquad (4\text{-}19)$$

where Θ is the minimum SINR expected at the UE to achieve the targeted throughput. E_I represents the expectation over the interference at i^{th} UE. η represents the noise power at the receiver and I is the interference observed at i^{th} UE is $\sum_{j=1,j\neq i}^{N} r_{ij}$.

As the interference is independent to the received power and the interferers are independent as well, the probability of coverage p_c can be derived as

$$p_c = exp\left(\frac{-\Theta\eta {d_{ij}}^{\alpha}}{P_i}\right)\prod_{j=1,j\neq i}^{K}\frac{1}{1+\Theta\frac{P_j}{P_i}\frac{{d_{ii}}^{\alpha}}{{d_{ij}}^{\alpha}}} \qquad (4\text{-}20)$$

4.4.2.3 Requirements

The targeted goals for COHERENT using DAS, throughput improvement and interference coordination, are completely lower layers oriented, i.e. MAC and PHY layers. The system mainly consists of distributed RRHs connected to a Central Unit (CU) for centralized processing of baseband signals. The usual communication between the CU and RRHs uses typical baseband communication standards like common public radio interface (CPRI). The requirements at lower layers can be broadly classified into monitoring and actions. Based on the targeted goals, an optimized Network Graph (NG) will be generated at the C3 / RTC, which would be processed by C3 / RTC for controlling the DAS network. The necessary parameters and abstraction metrics will be provided by the MAC and PHY layers towards the C3/RTC.

4.4.2.4 Generation and usage of Network graphs

For efficient control and coordination DAS requires UE's distance and pathloss details w.r.t. relevant RRHs, besides intra and inter-cell interference information, scheduling information etc. Network graphs provide graphical representation of the required information for RTC / C3 in conjunction with the CU. The necessary parameters and abstraction metrics will be provided by the MAC and PHY layers towards the C3/RTC. C3 / RTC would in return provide the information like interference level on different RBs due to inter-cell interference, coverage information etc. back to MAC for efficient scheduling. In this way, the requirements will start from the lower layers and will flow towards RTC / C3 and back. The network graph will be generated by the RTC / C3 with the interference measurement information collected from each eNodeBs. The RBs affected by the cell edge users due to severe interference will be identified and a decision to arbitrate the RBs reservation for the cell edge users will be provided by C3 / RTC. In this regard, the SINR / CQI reports of the UEs of both the cells will be collected at the C3 / RTC to coordinate the interference.

Figure 4-56 is a block diagram representation of the processing blocks for DAS using the COHERENT framework. As the goals are MAC and PHY layers targeted, the necessary measurement parameters from these layers will be collected and sent at the eNodeB to RTC. The network graph (NG) formulation is depicted in

Figure 4-56 as a separate module, as it is mere an implementation choice. RTC is required for the case of real time information exchange, as in the case of resource scheduling. The blocks to the left of RTC are technology or RAT dependent. The RTC may send the information further to NIF function block, where technology independent / RAT agnostic abstractions, depending on the need, may be derived. C3 module interfaces the information from different eNodeBs or APs. The case described in this figure, is a scenario where the C3 interfaces an LTE eNodeB and a Wi-Fi AP.

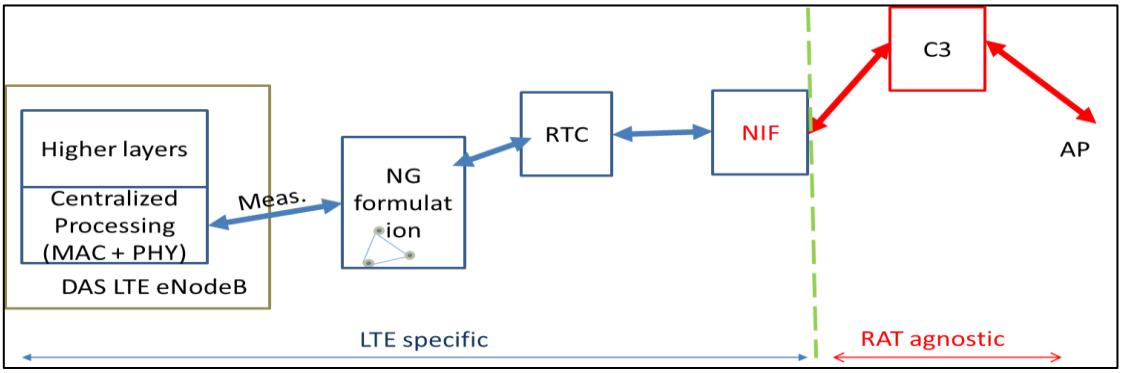


Figure 4-56 Block diagram representation of processing blocks for DAS in COHERENT

4.4.2.5 Abstracted parameters

For the improvement of per-user throughput using pairing and ICIC: the channel quality indicator (CQI) of each users, signal to noise plus interference ratio (SINR), scheduling information of each user, HARQ retransmissions rate of each user etc., can efficiently characterize the coverage problems or UEs suffering with high interference in the network. The following tables represent the set of abstraction parameters exposed and controlled for selection transmission based LTE-DAS.

Table 4-27 Exposed parameters

Parameter	Description	Direction	Deployment
P_c	Probability of coverage	UL/DL	RTC
$SINR_{i,j}$	Signal to interference plus noise ratio at UE w.r.t. all RRHs	UL/DL	C3 and RTC
CQI	Channel Quality Indication	DL	RTC, UER-RTC
UE radial distance	Radial distance of UE w.r.t. the center of cell	DL/UL	RTC, UER-RTC
Transmission power	Transmission power at each RRH and UEs	UL/DL	RTC, UER-RTC
Pathloss	Pathloss	UL/DL	RTC, UER-RTC

Table 4-28 Controlled parameters

Parameter	Description	Direction	Deployment
(UE, RRHs)	Association of UE with different RRHs	UL/DL	RTC
N_{RB}	Number of resource blocks allocated to a link	UL/DL	RTC, UER-RTC
Transmission power	Transmission power at each RRH and UEs	UL/DL	RTC, UER-RTC

4.4.2.6 Algorithms in C3

This work employs a statistical metric called 'probability of coverage', which ensures based on the channel model, the probability of achieving the targeted throughput for each UE based on the SINR at the UE location. As described in section 4.4.2.7, the network graphs are generated based on the probability of coverage that offers reliability in making the decision of pairing UEs with RRHs. These network graphs can be formulated based on the exposed parameters given in Table

4-27. Based on the network graphs generated, for e.g. Figure 4-58 and Figure 4-59, for the UEs located quite within the cell, i.e., the UEs not affected by the inter-cell interference, the pairing of UEs with the RRH is made at the RTC and the decisions are in turn given back to the CU as given in Table 4-28. C3 will be involved in making the decisions of UE, RRH pairing in the UEs affected by inter-cell interference. Hence C3 may make use of the probability of coverage based network graphs for efficient sharing of resources in order to manage inter-cell interference as in inter-cell interference coordination.

4.4.2.7 Results

One of the important parameters in selection transmission based UE and RRH pairing is monitoring the probability of coverage P_c of the UE w.r.t. all the RRHs based on the $SINR_{i,j}$ values at the UE locations. In section 4.4.2.2, the probability of coverage metric is derived, which is verified with an empirical model using Rayleigh distribution and Monte-Carlo simulations in MATLAB and the results match exactly. A single cell of 20 meters radius is considered for analysis. Within the cell, different RRH distributions i.e., the number of RRHs per cell is considered. The number of RRHs per cell has been selected to be 7, 19, 37, 61 and 91 based on the optimum circular packing of RRHs in the hexagonal area. For e.g. in Figure 4-57, a DAS system with 7 RRHs is being represented, with each RRH covering approx. 6 meters radius, different UE locations within the coverage of RRH_0 having radial distances of multiple of 1/10th of RRH coverage radius (approx. 0.6 meters) from the center is shown. Different UE positions as represented with 'x' and a down-pointed triangle represents the RRHs, numbered starting with 0. Equal transmit power from all the RRHs with Rayleigh fading channel and a noise floor of about -100 dB is assumed for the simulations. In Figure 4-58, a network graph is generated for the DAS scenario with 7 RRHs. The network graph describes the probability of coverage w.r.t. all RRHs for a target SINR of 20 dB at the position of UE_0, which is allocated at about 1/10th of RRH coverage radius. The thickness of the edge proportional to the value of probability of coverage can be depicted in the network graph. Hence in Figure 4-58, it is evident that pairing of UE_0 with

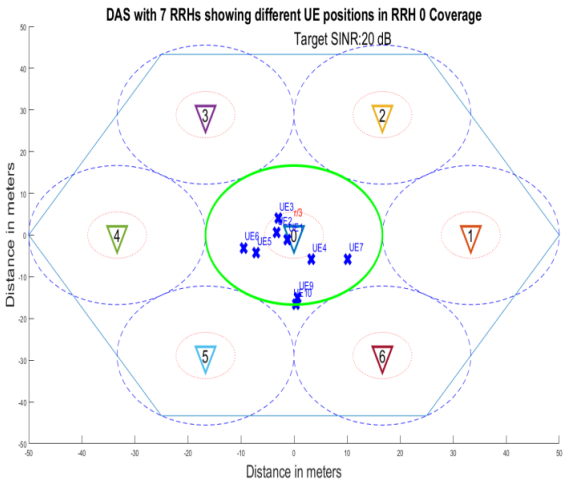


Figure 4-57 DAS system with 7 RRHs with various UEs

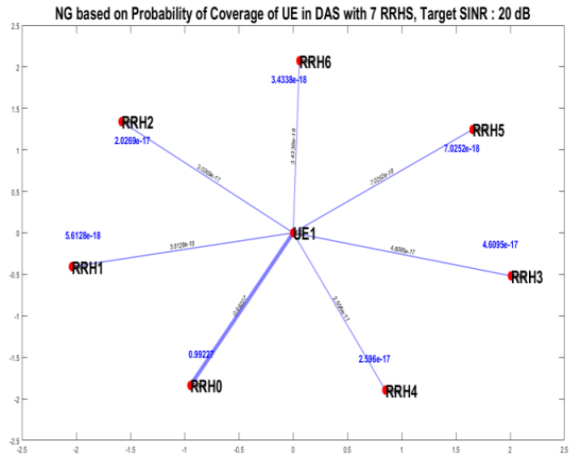


Figure 4-58 Network graph showing the probability of coverage at UE located at 1/10 distance of RRH_0 coverage radius for a Target SINR of 20 dB

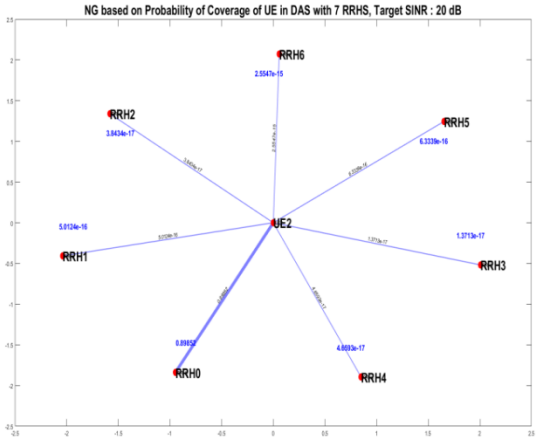


Figure 4-59 Network graph showing the probability of coverage at UE located at 2/10 distance of RRH_0 coverage radius for a Target SINR of 20 dB

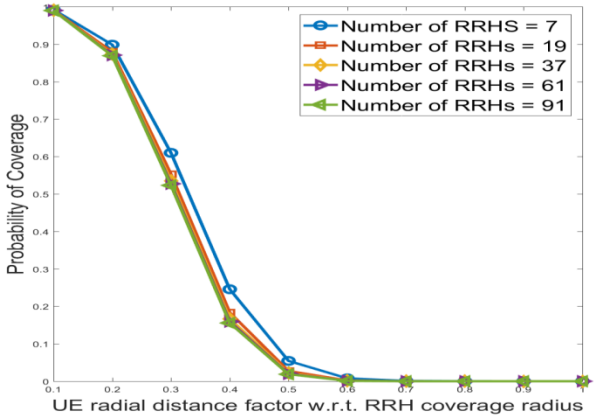


Figure 4-60 Probability of coverage for a targeted SINR of 20 dB in varied UE positions and the number of RRHs

RRH_0 has the best probability of coverage at UE_0, depicted by a thick line. Similarly depicts the network graph at UE_1, i.e., at a radial distance of 2/10th RRH radius from RRH_0.

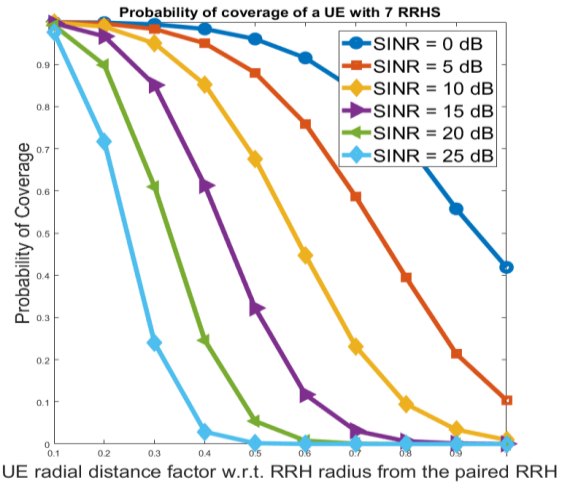


Figure 4-61 Probability of coverage for a targeted SINR of 20 dB in varied UE positions and the number of RRHs

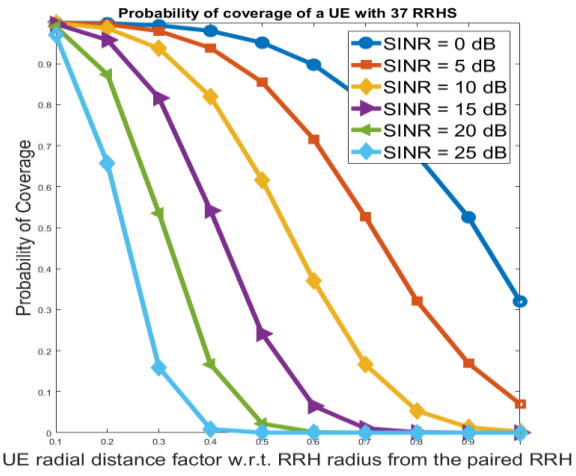


Figure 4-62 Probability of Coverage for different targeted SINR in a cell with 37 RRHs in varied UE positions

In Figure 4-60, the probability of coverage for a targeted SINR of 20 dB is simulated for different DAS scenarios. In this scenario, the number of RRHs per cell is varied to observe its effect on the probability of coverage for a targeted SINR. The curves show that there is a slight degradation in the probability of coverage if the number of RRHs in the same area is increased, but the more the number of RRHs the more probable will be the reutilization of time-frequency resources in the same area. Therefore, the advantage of having more RRHs outweighs the slight degradation. Figure 4-61 and Figure 4-62 show the probability of coverage w.r.t. different expected SINR, i.e.,

0, 5, 10, 20 and 25 dB, in case of 7 RRHs and 37 RRHs respectively. From these curves, depending on the targeted SINR, i.e., the targeted throughput the probability of coverage decreases rapidly if the UE is distant than 1/3 of the RRH radius from the location of RRH.

4.4.2.8 Conclusion

In this section, we have described a DAS system model and derived the probability of coverage metric for analyzing different DAS scenarios i.e., varied number of RRHs in a cell and targeted SINR at different UE positions. Probability of coverage is the employed abstracted metric for network graph generation, which ensures that the required throughput at the UE location is achieved statistically based on the channel model. Selection transmission based on pairing UEs with RRHs in considered as the DAS technique in order to efficiently pair the UEs to RRHs so that per-user-throughput and the system throughput are improved besides ensuring the reliability. Different RRH packings i.e. with 7, 19, 37, 61 and 91 RRHs per cell scenarios are presented. The probability of coverage at different UE positions based on targeted SINR, which is directly related to the required throughput at the UE position, in different DAS scenarios is analyzed. From the presented results we conclude that, depending on the targeted SINR, i.e., the targeted throughput the probability of coverage decreases rapidly at the UE positions greater than 1/3 of the RRH coverage radius. However, compared to the traditional multiple access in LTE standard, where non-overlapping scheduling of resources is defined, one can make use of the probability of coverage at the UE position, throughput requirement of the UE in order to exploit the opportunities to schedule overlapping resources. Hence pairing RRHs with UEs enables to find the opportunities within the cell to re-use the RBs, which can significantly improve the spectral efficiency. Therefore DAS provides viable solutions for the future 5G mobile communication systems.

4.4.3 Massive MIMO Exploitation to Reduce HARQ Delay in PHY/MAC

4.4.3.1 Motivation and goals

As recommended by 5th generation Public-Private Partnership (5G PPP) and Next Generation Mobile Networks (NGMN), one of the most important 5G requirements is to minimize the delay within the network for delay-sensitive services (like Voice Over IP, video telephony, real-time gaming and M2M) to achieve low and ultra-low latency [59] [60]. With the significant reduction of delay in the core of mobile networks due to structure simplification (e.g. System Architecture Evolution (SAE) in 3GPP LTE system), the delay reduction in the wireless part is still a challenge. Hybrid Automatic Repeat Query (HARQ) schemes cause the major part of this delay in particular when the links, on which losses occur, have long path delay [60].

The main objective of this work [61] is to exploit massive MIMO technology for reducing HARQ retransmission delay; then to define the metrics and parameters that are required to generate the network graphs in C3/RTC; and finally to use them for applying the proposed resource allocation algorithms in C3. Using of beamforming in massive MIMO reduces the bit error rate, and that reflects directly on reducing the maximum required number of HARQ retransmissions. Such exploitation can be applied in both LTE and WiFi and also in multi-RAT networks with the use of COHERENT architecture. This application is directly linked to the RAT sharing use case (UC2.MM) described in D2.1 [2].

4.4.3.2 Requirements

In 3GPP technical report for Universal Terrestrial Radio Access Network (UTRAN) [62], an estimation of 5 msec retransmission time is used in user-plane. Figure 4-63 shows the delay components of LTE Frequency-Division Duplexing (FDD) in user-plane [62], which is related to the expected number of packet transmission N and it can be expressed in msec as:

$$T(N) = 1 + 1.5 + 1 + 5(N-1) = 3.5 + 5(N-1) \quad (4\text{-}21)$$

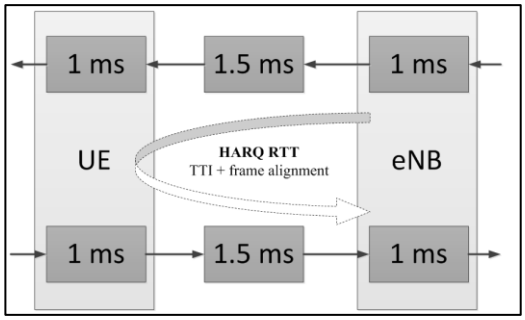


Figure 4-63 User-plane delay components of LTE FDD in 3GPP standards

Thus, the standards limit the maximum number of packet retransmissions to control the maximum delay. For example, normally in 802.11 WiFi system the number of maximum allowed retransmissions is L = 7 (for short packets), meanwhile some literature suggest to limit it for LTE in the downlink to save the radio resources for user-equipment (UEs) with poor radio link [63].

4.4.3.3 Generation and usage of Network graphs

Each node at the graph represents a massive MIMO base station (LTE or WiFi) or user equipment. The edges between the nodes are metrics that include the angle of arrival (AoA), the signal to interference plus noise ratio (SINR), the bit error probability P_b, and the probability of packet retransmission P_{ret} which can be calculated based on knowing P_b, the packet size, and the channel coding parameters. In general these metrics are technology-agnostic and allow the controller to give higher priority for delay-sensitive applications and to allocate resources from the RAN that guarantee less delay.

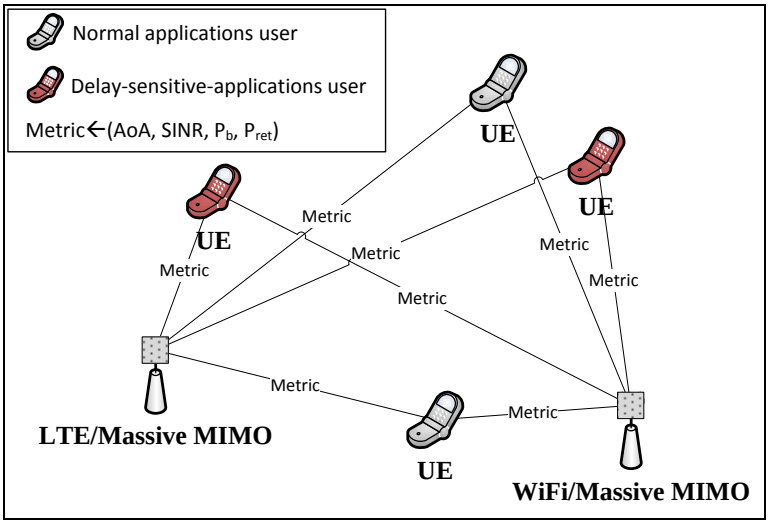


Figure 4-64 Network graph example of HARQ delay reduction by massive MIMO

4.4.3.4 Abstracted parameters

In this work, the abstracted measurements and parameters are: the angle of arrival/departure (AoA/AoD) of the DL/UL users that are using delay-sensitive services, also the probability of packet retransmission P_{ret} based on the SINR and probability of bit error rate P_b.

Table 4-29 Exposed and controlled parameters

Parameter	Description	Direction	Deployment
AoA	Angle of arrival in massive MIMO array at the receiver	UL	RTC,C3
AoD	Angle of departure in massive MIMO array at the transmitter	DL	RTC,C3
P_{ret}	The probability of packet retransmission	UL/DL	RTC,C3
P_b	The bit error probability	UL/DL	RTC,C3
SINR	Signal to Interference plus Noise Ratio	UL/DL	RTC,C3, UE
Controlled parameters			
Serving Cell	Serving RAN Node of end device	UL/DL	RTC, C3

4.4.3.5 Algorithms in C3

Massive MIMO indoor environment is created by a 3D ray-tracing tool to simulate the received power, phase, delay, and angle of arrival of each ray in the channel. Two sizes of rectangular antenna array (10x10 and 7x7) are used to evaluate the performance of beamforming for enhancing the bit error rate. This reflects directly on the mean and maximum numbers of packet retransmission. The angle between user-equipment $\Delta\emptyset$ is considered to evaluate the effect of the interference between beams on HARQ.

The wireless system model and the indoor simulation environment are shown in Figure 4-65:

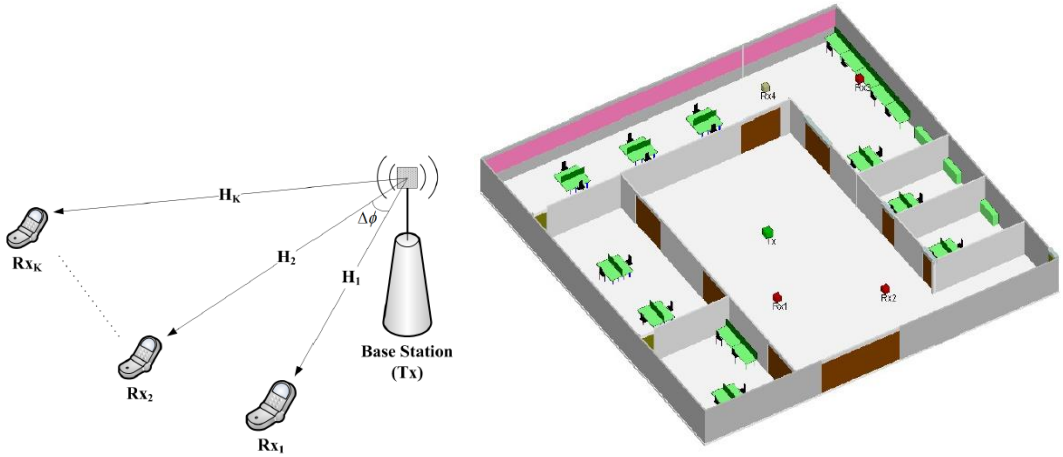


Figure 4-65 The wireless system model and the indoor simulation environment for massive MIMO

In general HARQ delay can be written in terms of N the mean number of transmission for a packet:

$$T(N) = (N-1)T_u + T_s \qquad (4\text{-}22)$$

$$N = \frac{1}{1-P_{ret}} = \frac{1}{\sum_{i=0}^{t}\binom{n}{i}P_b^i(1-P_b)^{n-i}} \qquad (4\text{-}23)$$

where T_s and T_u are the time for each successful and unsuccessful transmission attempts respectively, *n* is the packet size, *t* is the correction capability of the channel coding.

The maximum delay is when the maximum allowed number of retransmission is reached.

$$T_L = (L-1)T_u + T_s \qquad (4\text{-}24)$$

Meanwhile the probability of maximum delay $Prob\{T_L\}$ is the probability of L retransmission in terms of bit error probability:

$$Prob\{T_L\} = (P_{ret})^L = [1 - \sum_{i=0}^{t}\binom{n}{i}P_b^i(1-P_b)^{n-i}]^L \qquad (4\text{-}25)$$

where P_b can be obtained from simulation with changing the *SNR* and the angle between the two users $\Delta\emptyset$ to consider the multi-user interference.

The simulation results focus on the number of retransmissions of a packet when the following variables are changing:

- SNR: Signal to Noise Ration in wireless channel
- $\Delta\emptyset$: Angle between the two user-equipments
- Size of antenna array: 100 (10x10) or 49 (7x7)

The simulation results are averaged over 1000 channel response simulations.

Shared coverage area (cells borders) scenario

One example scenario is to allocate the users of two cells in the shared coverage area (either on cells edges or in case of heterogeneous networks) for minimizing the interference between users. In this mutli-cell scenario, centralized coordination by C3/RTC is required as follow:

- The base stations send the metrics to C3 in order to generate the required network graph for this application.
- Using the generated network graph, and by dividing the delay-sensitive users into two groups as in Figure 4-66, C3 forms two scenarios of the users allocation (Group A belongs to BS1, group B belongs to BS2 and vice versa).
- Evaluating the two scenarios based on the expected performance for each one (maximizing the sum of angle separation and minimizing the sum of inter-user interference), then selecting the best scenario, meanwhile the other scenario can be considered as the second best scenario in comparison with no no-coordination scenario without using the algorithm. Nevertheless both scenarios will result a smaller number of HARQ retransmissions and thus a smaller delay as shown in the results compared to no-coordination scenario.

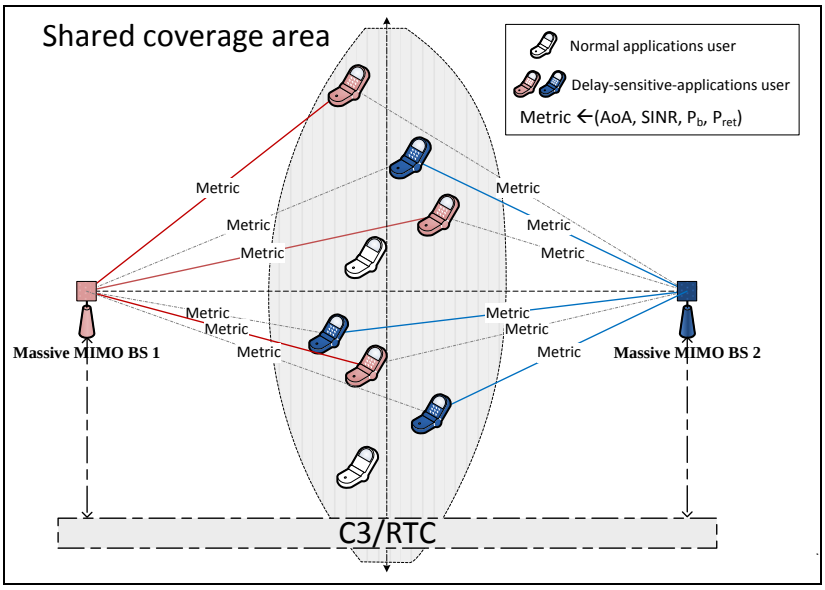


Figure 4-66 Exploit massive MIMO to reduce HARQ delay for users on the cells borders

4.4.3.6 Results

Figure 4-67 shows the mean number of HARQ retransmissions based on the achieved bit error rate by the simulation, in comparison with the theoretical analysis which is done in (4-23). The channel coding scheme BCH (Bose-Chaudhuri-Hocquenghem-Codes) is used with n = 63, t = 13.

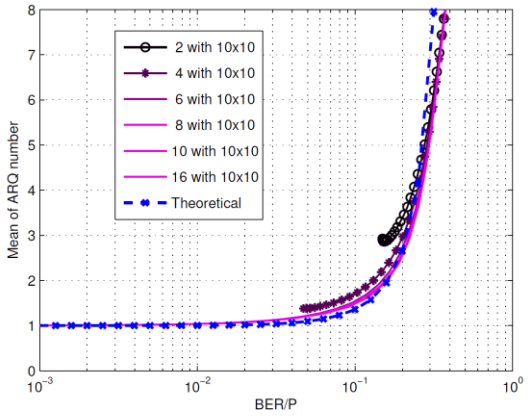


Figure 4-67 The mean number of HARQ retransmissions with variant ΔØ

Figure 4-68 shows the mean and maximum number of retransmissions respectively considering L=10. Two array sizes are considered 7x7 and 10x10 and with changing ΔØ from 2° to 14°.

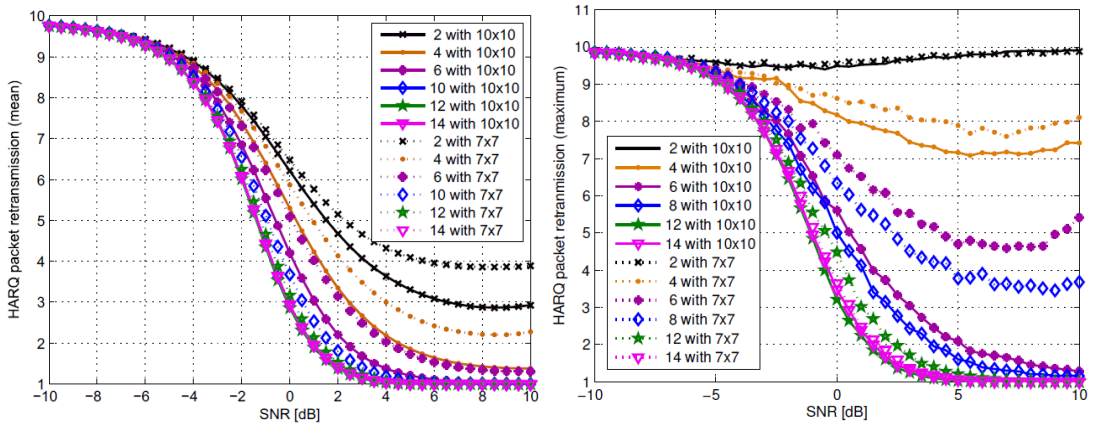


Figure 4-68 The mean and maximum numbers of HARQ packet retransmission with variant ΔØ

The figures show that the more separation between users (higher ΔØ) the less interference and hence the less number of retransmissions. Better performance of a larger antenna array is expected as it reduces the interference between users.

Figure 4-69 depicts the results of the proposed algorithm simulation in C3 for users on cell borders.

The shared coverage area on cell borders is assumed as a strip with a width of 10% of the total distance between the two base stations in Figure 4-66. The results are averaged over 1000 simulation, each one with a variant number of users between 6 and 10 users in random angles and positions in the shared coverage area.

The left figure shows the average BER over all users in the three scenarios (best, 2nd best, no coordination). It is clear that the algorithm can enhance the system performance considerably.

The right figure show the maximum number of retransmissions averaged over all users. This value directly refers to the HARQ delay. The algorithm can reduce HARQ delay to less than 50% compared to no coordination scenario.

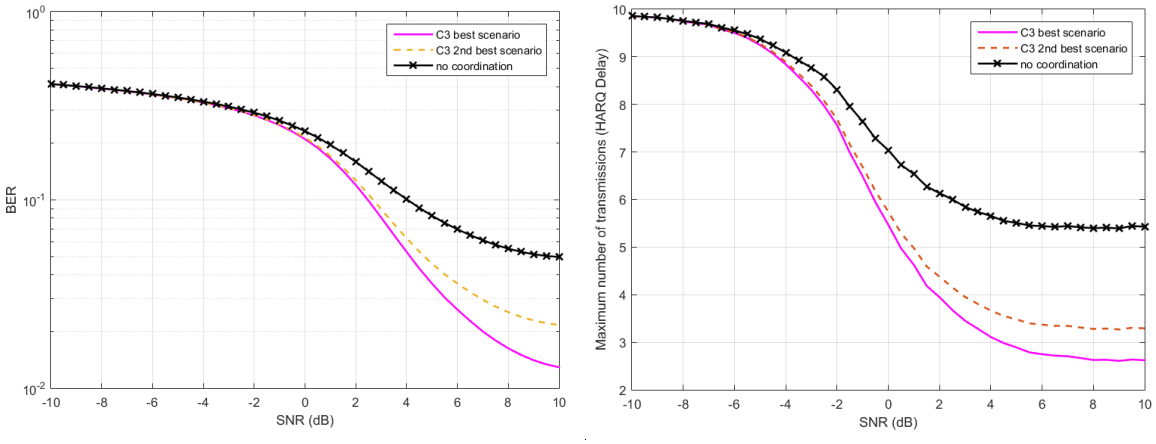


Figure 4-69 The averaged BER and the maximum numbers of HARQ retransmissions with the proposed algorithm in C3

4.4.3.7 Conclusion

The aim of this work was to show how the abstraction and network graph can be used for new technologies like massive MIMO within COHERENT architecture and for heterogeneous mobile networks. The results of this work show that exploiting this new technology together with a centralized coordinator will allow better performance for delay sensitive applications which were mentioned in 5G recommendations. Based on such results, smarter resource allocation algorithms in C3 and RTC can be developed to maximize the sum on separation angles and to minimize the interference between users who have the highest priority in terms of delay in single-RAT or multi-RAT systems. However, in spite of the achieved reduction amount of the delay in PHY/MAC, further investigations have to be done to evaluate the delay generated by applying the algorithms and performing the coordination in C3/RTC in order to define more realistic results. So in general some small loss in this reduction is expected in implementation.

4.4.4 Inter-RAT spectrum resources sharing

4.4.4.1 Motivation and goals

The main goal of this activity is to provide abstraction and control background for management of limited spectrum resources in a dense heterogeneous network. The considered scenario is a coexistence of at least two different Radio Access Technologies (e.g., Wi-Fi and LTE or LTE and GSM) in the same spectrum. It is a challenging scenario as interference between various systems can be very high. It has to be calculated and considered while allocating resources to various users utilizing various RATs. Provision of such functionality allows for increased spectral efficiency and e.g. smooth migration from older RAT technology to newer one utilizing a single frequency band.

Although there exists some long or medium term spectrum sharing schemes, e.g., Licensed Shared Access described in WP4, these do not allow adjusting dynamically both systems to instantaneous traffic demand in both systems. Below, an inter-RAT scheduling algorithm is presented. It has been disseminated by a paper [90]. This application is in line with Use Case 2 Flexible spectrum access defined in D2.1 [2].

4.4.4.2 Requirements

The proposed algorithm assumes both controlled systems utilize TDMA/FDMA with possibly different minimum time and frequency allocation block (granularity). Random access schemes (e.g., CSMA) are typically utilized in systems providing ”Best Effort” traffic, e.g., WiFi. In these systems inter-RAT interference is managed typically in a statistical fashion, e.g., by varying back-off window duration in WiFi/LTE-U coexistence.

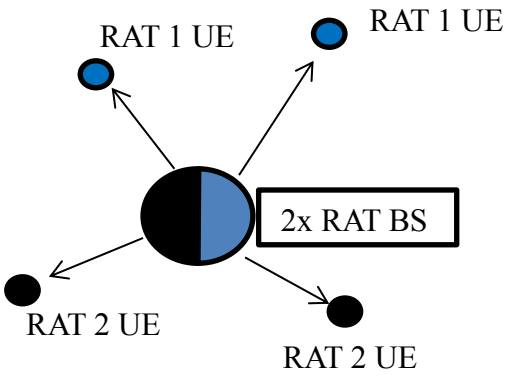


Figure 4-70 Scenario of inter-RAT scheduling

In order to provide tight time/frequency alignment (e.g., common clock) between both RATs it is assumed that both Base Stations (BSs) are collocated in space. There can be two separate BSs mounted at the same mast/connected to the same antenna or a single, dual RAT BS. This scenario is graphically presented in Figure 4-70. Each user has to report its channel conditions (common requirement for modern communications systems). Additionally, interference coupling between each resources chunk in one system and any resources chunk in the other system are to be known. This is required in order not to allow one transmission to significantly deteriorate the other scheduled transmission quality.

It is foreseen that C3 cannot provide short term resources allocation because of control information delay. Short time resources allocation can be carried by real-time controller implemented close to the transmission points. The C3 would require some long or medium-term statistics on e.g., users' throughput requirements and achieved rates, load in adjacent cells etc. This can allow for tuning the implemented scheduling algorithm, e.g., by means of γ parameter that specifies averaging window length of historical rate (details provided in [90]). Indirectly, it allows balancing total throughput and fairness between both RAT users.

4.4.4.3 Abstracted parameters

The network graph can be created locally at RTC. In the proposal, the nodes denote a given transmitter utilizing given resources block in a given timeslot. This is connected to the receiver (additional nodes) with the edge weight denoting achievable Modulation & Coding Scheme (MCS). At the same time, this allocation results in interference to many other transmissions in many other Resources Chunks of the other RAT. This can span many resources chunks in frequency and in time (past, current and next timeslots). This can be denoted as a graph edges with their weights being Adjacent Channel Interference Ratio (ACIR- metric based on Adjacent Channel Leakage Ratio describing transmitter and Adjacent Channel Selectivity describing receiver).

The scheduling algorithm needs parameters listed in Table 4-27 to be able to maximize normalized rate, i.e., ratio of currently available rate to the historical rate and maintain acceptable interference to previously allocated resources. While CQI can be directly related to the possible rate (RAT-specific mapping function required) in a given link, RSSI, RSRP and 'Interference Margin' allow to calculate how strong the interference from the other RAT can be in order to keep the same CQI value (this is one of possible approaches). Next, the ACLR and ACS values allow to calculate amount of interference caused by previously allocated resources, and by the currently resource to the other resources. In [66] an analytical approach has been provided to calculate interference coupling between, e.g., GSM-type transmission and OFDM. The parameter γ can be controlled at C3 in order to find balance between fairness and maximizing summarized rate between both RATs.

Table 4-30 Exposed parameters

Parameter	Description	Type	Direction	Deployment
CQI	Channel Quality Indicator per resource block	Complex	DL/UL	RTC
RSRP	Received Signal Received Power- received 'wanted' signal power	Simple	DL/UL	RTC
Interference margin	How can the SINR be decreased to still operate at the same CQI? In dB	Complex	DL/UL	RTC
ACLR	Adjacent Channel Leakage Ratio, per each resource chunk (in time and frequency) that will be affected by the allocation.	Complex	DL/UL	RTC
ACS	Adjacent Channel Selectivity, per each	Complex	DL/UL	RTC

	resource chunk (in time and frequency) that will be affected by the allocation.			
RSSI	Received Signal Strength Indicator- total power of received signal	Simple	DL/UL	RTC

Table 4-31 Controlled parameters

Parameter	Description	Type	Direction	Deployment
γ	Coefficient used in historical UE rate averaging. In range (0,1).	Simple	DL/UL	C3

4.4.4.4 Algorithms in C3

The block diagram of the proposed scheduling algorithm is provided in Figure 4-71. In the picture the parts specific for a given RAT are plotted in blue or black boxes. The RAN-agnostic parts (although having to use some data obtained from each RAT) are given in white boxes. This algorithm is based on a state-of-art, proportional fair (PF) algorithm proposed for LTE system in [67]. The choice of PF solution is reasonable as it takes achievable rate into account (based on the current CQI value). Moreover, it promotes the users that had low past throughput (calculated using γ weighting factor). If max-rate algorithm is used and one RAT is more spectrally efficient than another, it will obtain all the resources.

The proposed solution takes into account maximal rate achievable by a given user. It is divided by its historical rate. Users are allocated according to the order of these normalized rates. Each user is allocated all the resources maximizing its throughput out of the maximum available. However, before each allocation is fixed, its interference to all previously allocated users has to be checked. If it does not exceed a given threshold (defined by 'Interference margin') the allocation is confirmed. This "linear" procedure reduces computational complexity. After distributing all resources or checking that all users are scheduled, the procedure ends. The calculated sets of users, allocated resources chunks and calculated Modulation-Coding Schemes are sent to both RATs.

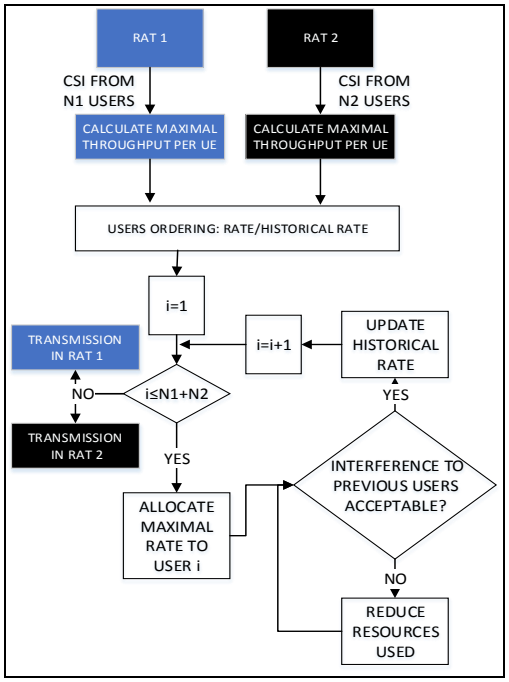


Figure 4-71 Block diagram of the proposed inter-RAT scheduling

4.4.4.5 Results

The algorithm was tested for spectrum sharing between LTE and GPRS technologies. These systems operate at carrier frequency 900 MHz with bandwidths 1.4 MHz (6 RBs) and 200 kHz for LTE and GPRS, respectively. Both systems utilize mean transmit power of 43 dBm. Initially, there are 16 LTE and 4 GPRS users equally distributed in a round cell of radius 2000 m. The pathloss is calculated according to 3GPP TR 36.942. The multipath fading is characterized by Extended Vehicular A model with UEs speed of 5 km/h. While the interference coupling is calculated as provided in [66], the noise floor is assumed to be -165 dBm/Hz. Base station antenna gain equals 15 dBi, and isotropic antenna is assumed for UEs. For GPRS, 4 MCS are used with their parameters taken from [64]. For LTE 15 MCS are considered with parameters defined in [65][64].

The simulation was carried over 1000 random users' positions (pathloss and fading change). In a single position 100 ms of continuous fading change was simulated. The proposed Multi-RAT scheduler has been compared with separate LTE and GPRS schedulers operating in the same band.

The results in Figure 4-72 show CDF of mean UE rate for γ=0.01. It is visible that in both LTE and GPRS systems the Multi-RAT scheduler outperforms two separate schedulers in terms of achievable rate. It is visible both in terms of mean UE rate and $10^{th}/90^{th}$ percentile users' rate. This is a result of interference forecasting. While "separate schedulers" treat inter-RAT interference as decreased SINR and utilize channel with lower MCS, the Multi-RAT scheduler refrains from allocating significantly overlapping (in terms of interference) resources. A resources chunk of high quality can be used with higher MCS.

This effect is visible in Figure 4-73, where an example of time-frequency grid utilization is provided. While separate schedulers utilize all available resources, Multi-RAT schedulers e.g., limits LTE transmission over RBs overlapping with GSM band when GPRS timeslots are heavily used.

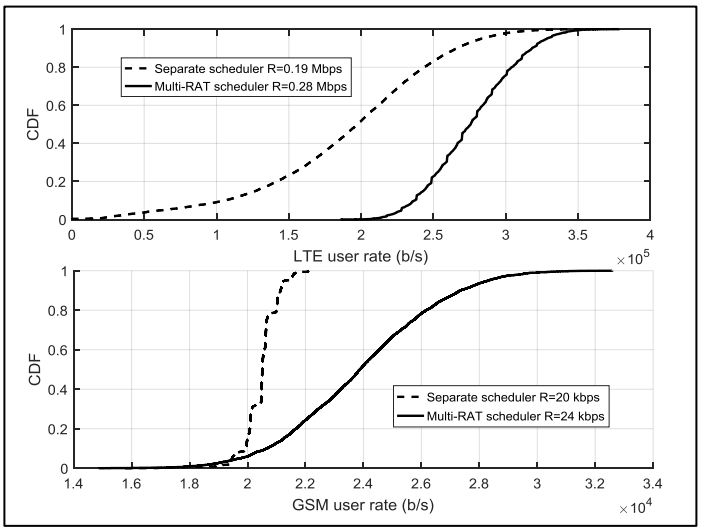


Figure 4-72 Cumulative Distribution Functions of mean LTE/GPRS UE rate for Multi-RAT scheduler and separate schedulers.

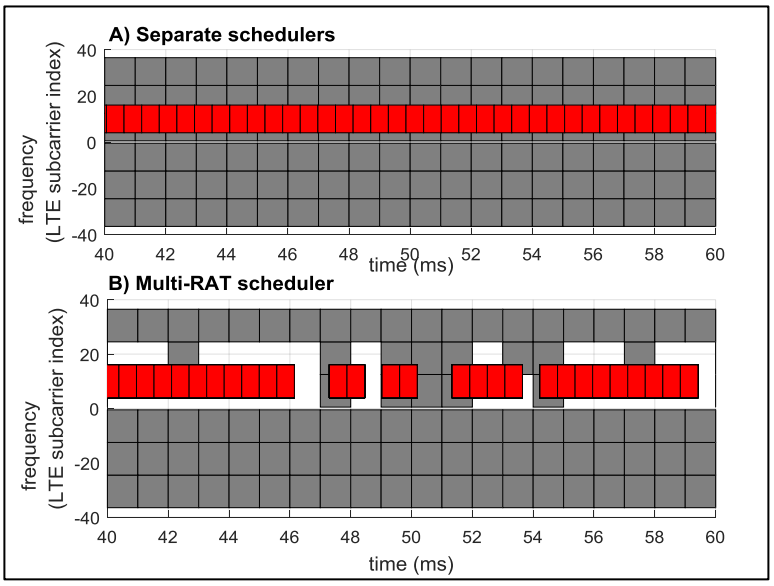


Figure 4-73 Example of time-frequency grid utilization (red rectangle: utilized GPRS timeslot, grey rectangle: utilized LTE resources block)

An important feature of the proposed multi-RAT scheduler is balancing traffic depending on changing demand in both RATs. In Figure 4-74 mean LTE/GSM BS rate is shown as a function of GSM UEs number. Total number of UEs in both systems is fixed to 20. It is visible that Multi-RAT scheduling provides more rate to a RAT whit more users requiring traffic. On the other hand, standard schedulers do not have this property. The rise in mean GSM BS rate for 18-19 GSM UEs and "Separate schedulers" is a result of low number of LTE UEs. These are not utilizing low-SINR RBs collocated with GSM transmission. GPRS transmission increases its rate.

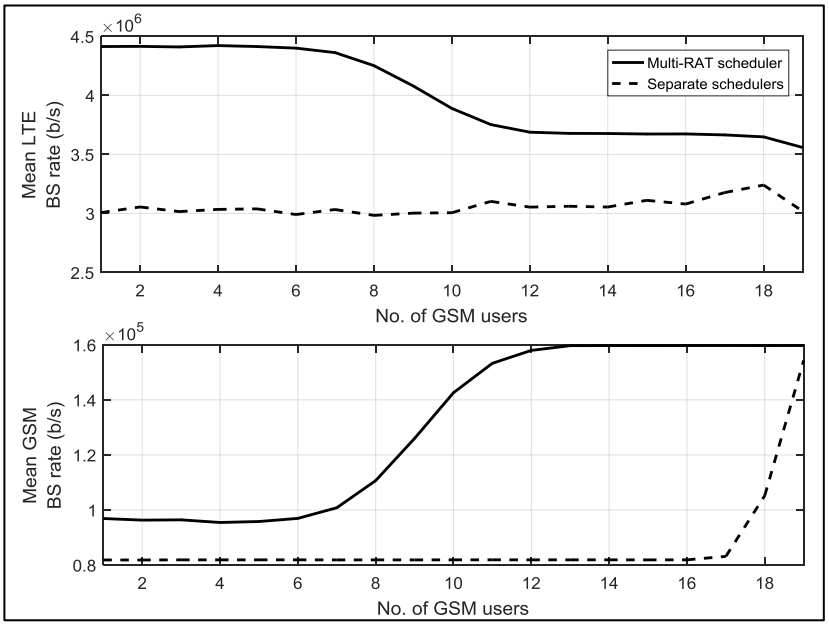


Figure 4-74 Mean LTE/GSM BS rate as a function of UEs number.

4.4.4.6 Conclusion

The proposed inter-RAT scheduling algorithm allows to share limited time-frequency resources among users utilizing various RATs. The proposal can be of high importance in dense heterogeneous network utilizing limited frequency resources.

4.4.5 Inter-Cell Interference Coordination

4.4.5.1 Motivation and goals

HetNets with related Inter-Cell Interference Coordination (ICIC), need coordination of the RAN nodes in multiple tiers. We aim to solve disputes for resources in HetNets implementing ICIC at network-level. We propose a logical description of the RAN by Logical RAN Entities (LREs), under the control of the C3, where each LRE is a set of cells in the HetNet. The state of the RAN is abstracted, and described by a network graph, with nodes representing LREs. We consider a HetNet composed by macro and pico cells, with CRE, transmitting in downlink.

4.4.5.2 Requirements

A centralized control framework with a C3 is required for coordination. The users are responsible for measuring RSRP, identifying dominant interferences, and reporting a metric based on aggregated spectral efficiencies, with and without the dominant interferences. The RAN nodes are responsible for creating LREs, aggregating user estimates, and reporting parameters to the C3. The C3 is in charge of constructing the network graph, and implementing a suitable logic for the coordination of the network.

4.4.5.3 Generation and usage of Network graphs

A network graph is constructed to manage the interference situation produced by the strong interference that macro BSs exert into small cell users in the CRE region. LREs can be created and removed dynamically, for example, after cell-user association. Each user associated with a macro cell *m* is further associated with a LRE (m). Each user associated with a small cell *s*, and with the presence of a set M of macro cell strong interferers, is associated with the LRE (s, M). If M is the empty set, then we denote (s, M) by (s), and if M = {*m*} is a singleton, then we denote (s, M) by (s, m).

RAN network entities compute a utility for users, and a sum utility for LREs, from the spectral efficiencies reported by the users (see [91] for details). An adjacency matrix is created, which describes conflicts between LREs, and then it is used to construct the network graph. Edges in the graph are established between LREs that have at least one common member. Figure 4-75 shows an example of a HetNet with three macro cells, two small cells, and associated users. An example of a network graph of logical entities derived for the HetNet is shown in the figure.

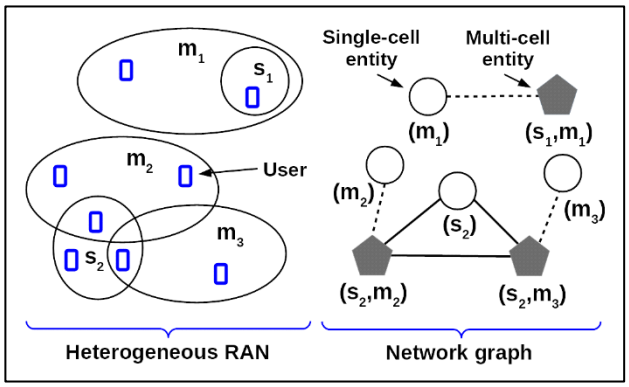


Figure 4-75 Example of a HetNet and network graph of logical entities

Resources are interchangeable among LREs respecting conflicts. In each macro cell, resources can be muted with an independent muting ratio. Resources for local scheduling in each LRE are assumed infinitely sub-divisible among the users in the LRE. The resource allocation problem is a multi-coloring problem of the graph. A Tabu Search algorithm (TS-NG-ICIC) is implemented, to find optimal muting ratios through movements in the graph, subject to the constraints discussed in [91]. The muting ratios are communicated to the cells. The RTCs determine locally the scheduling weights for the users, constrained to the muting ratios stated by the C3.

4.4.5.4 Abstracted parameters

Exposed parameters are the number of users and the aggregated spectral efficiencies in each LRE, and the list of LREs in each cell. Controlled parameters are the muting ratios for macro cells. The following tables summarize the abstracted parameters, exposed and controlled, for ICIC.

Table 4-32 Common exposed parameters

Parameter	Description	Direction	Deployment
n_{LRE}	Number of users in a LRE. .	DL	RTC
SE_{LRE}	Aggregation of spectral efficiencies in a LRE.	DL	RTC
LRE	LREs in a cell	DL	RTC

Table 4-33 Common controlled parameters

Parameter	Description	Direction	Deployment
MR_m	Muting ratio for macro cell *m*	DL	C3

4.4.5.5 Algorithms in C3

The algorithm to be implemented in the C3 (TS-NG-ICIC) is explained in [91].

4.4.5.6 Results

We tested the algorithm in a HetNet as described in [91], performing Monte Carlo simulations. Figure 4-76 shows the simulation results for a HetNet with M = 12 macro cells, 48 small cells and U=720 users. Performance is evaluated in terms of the CDF of average user rate. We compare the performance of the algorithm TS-NG-ICIC against the cases of no muting and fixed muting in all

the network. We observed almost no gain in the system average user rate, but up to 77% gain per user in the 5th. percentile, compared to fixed muting (see [91] for details).

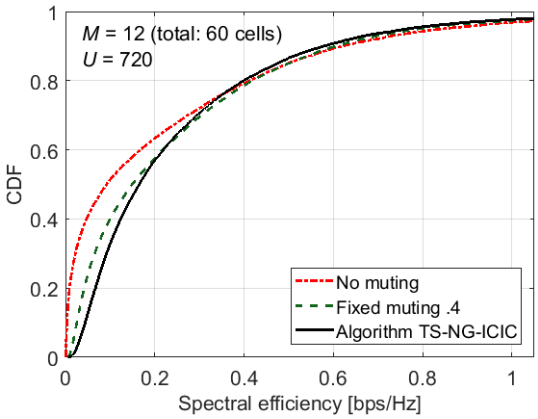


Figure 4-76 Simulation results for a HetNet with M=12 macro cells, 48 small cells and U=720 users.

4.4.5.7 Conclusion

We proposed a logical description of the RAN by Logical RAN Entities, and described the state of the RAN by a network graph, to solve disputes for resources in HetNets implementing ICIC at network-level. Results show the feasibility of performing network-level coordination with the proposed approach, and positive gains for users in the lower percentiles, without loss in system average user rate.

5 Control Interface APIs

5.1 Introduction

This section presents the procedures used to define the messages between C3 and the relevant network entities. The general approach for message definition follows the 3GPP approach as used in X2 interface defined in LTE standard TS 36.423 [87] and all the other interfaces defined by 3GPP RAN3, including interfaces to be used for 5G New Radio (NR) specifications.

3GPP RAN3 defines two classes of Elementary Procedures (EP):

- Class 1: Elementary Procedures with response (success and/or failure). Even if not specified explicitly, Class 1 procedures are REQUEST procedures.
- Class 2: Elementary Procedures without response (considered always successful). Even if not specified explicitly, Class 2 procedures are used mainly for Reports.

We consider that similar reports shall be provided in DL and UL, however for the sake of simplicity will be described mainly the reports for DL.

This section illustrates the coding for some selected messages in D2.2 [3] and this document and does not represent an exhaustive API for COHERENT SBi.

The material of this section and the relevant Annex has been used in contributions to 3GPP RAN3 on 5G new F1-C, F1-D and E1-C interfaces and to ETSI BRAN on Central Control of WiFi and LAA networks (draft TR 103 494).

5.2 Control Interface API for F1-C interface

In this section we will focus on procedures specific to COHERENT approach used over F1-C interface (CU-DU) and E1-C (between CU-C and CU-U) interface. The message detailed coding is included in Annex A.

5.2.1 Procedures and messages over F1 interface

Table 5-1 presents a summary of the procedures defined over F1-C (F1-Control) interface for transporting the new information defined in COHERENT.

Table 5-1 Class 1 Elementary Procedures Over F1 Interface

Elementary Procedure	**Initiating Message**	**Successful Outcome**	**Unsuccessful Outcome**
		Response message	**Response message**
Resource Status Reporting	RESOURCE STATUS REPORT REQUEST	RESOURCE STATUS RESPONSE	RESOURCE STATUS FAILURE
C3-innitiated action request	C3-INNITIATED ACTION REQUEST	C3-INNITIATED ACTION RESPONSE	C3-INNITIATED ACTION FAILURE
R-TP-initiated action request (event driven)	R-TP-INITIATED ACTION REQUEST	R-TP-INNITIATED ACTION RESPONSE	R-TP-INNITIATED ACTION FAILURE

The RESOURCE STATUS REPORT REQUEST is a message indicating what resources shall be reported and the way of reporting, i.e. periodic reporting, for which the time period for reporting is indicated as an IE, or request-driven reporting.

The semantics of the message is in Annex A,

Table 5-2 Class 2 Elementary Procedures Over F1 Interface

Elementary Procedure	Initiating Message
DU-triggered measurement report	DU-TRIGGERED MEASUREMENT REPORT

5.2.2 Message Functional Definition and Content

5.2.2.1 RESOURCE STATUS REPORTING

RESOURCE STATUS REQUEST

This message is sent by C3 (CU) to served R-TPs (DUs or gNBs or eNBs) to initiate the requested measurement according to the parameters given in the message.

Direction: C3 → R-TP.

The semantics of the message is described in section A.5.1.

RESOURCE STATUS RESPONSE

This message is sent by an R-TP to indicate that the requested measurement, for all or for a subset of the measurement objects included in the measurement request is successfully initiated.

Direction: R-TP → C3.

The semantics of the message is described in section A.5.2.

RESOURCE STATUS FAILURE

This message is sent by the R-TP to indicate that for none of the requested measurement objects the measurement can be initiated.

Direction: R-TP → C3. The semantics of the message is described in section A.5.3.

RESOURCE STATUS UPDATE

This message is sent by R-TP to C3 to report the results of the requested measurements.

Direction: R-TP → C3.

The semantics of the message is described in section A.5.4.

5.2.3 C3-INNITIATED ACTION

The messages included in this procedure are:

- C3-INNITIATED ACTION REQUEST
- C3-INNITIATED ACTION RESPONSE
- C3-INNITIATED ACTION FAILURE

Messages are introduced in section A.6.

The semantics is designed such to provide a generic message, where the name of a specific action can be selected from a list.

5.2.4 R-TP-INNITIATED ACTION REQUEST

Below are some of the possible R-TP (aka UD) requests:

- UE HANDOVER REQUEST
- R-TP LOAD BALACING

5.2.5 R-TP-INNITIATED ACTION RESPONSE

In case of HO or LOAD BALANCING, C3 provides the target list of new serving Cell(s) and their serving R-TPs to which the UE is hand-overed. R-TP will relay this information to UE.

5.2.6 R-TP-INNITIATED ACTION FAILURE

R-TP provides to C3 the cause of failure.

5.3 Control of heterogeneous IEEE 802.11– 3GPP LAA network

5.3.1 General Reports

An introduction to reports for C3 is provided in Annex A.4.1 .

5.3.2 Other reports and C3 Commands

An introduction to other reports and C3 commands is provided in Annex, section A.7.1.

5.4 Data User plane interaction with central coordinator

5.4.1 Transmission Point Selection

The C3 selects the Base stations, R-TP or R-TPs (in multi-connectivity) serving the UE based on a multitude of criteria, for example link quality, availability of time/frequency resources, backhauling delay, suitability to the network slices implemented by UE, etc.

5.4.2 Virtual Hybrid Routing Function

The last function in the DL chain shall be a routing function, providing bearer re-direction in case of hand-over, with the following functionality:

- Inserts in the IP header the IP address of the target R-TP and its TEID (Tunnel End ID)
- Sends a packet to C3 including the last LSN (Last Sequence Number) of each included Layer and the number of bytes in the UE specific queues, per UE and per QCI.

The initial semantics is provided in section A.8.

The first function in the UL chain shall be a DPI (Deep Packet Inspection) function, with the following functionality:

- Retrieve the source IP Address and the R-TP ID of the last correctly received packet
- Retrieve the LSN of each included Layer in the last received packet
- Send a packet to the C3 including the retrieved information; in this way the C3 will be informed about the status of transmission/reception.

For detecting a bottleneck in the backhaul network, the R-TP should also send the count of bytes in R-TP queues.

5.4.3 HARQ and ARQ

Another example of partition is related to handling of HARQ and ARQ processes; given the possible processing of RLC and MAC layers in VMs, the type of redundancy (how many retransmissions) can be provided by the C3, but the actual TB building, such to include in the MAC or RLC header the retransmission index, can be handled by R-TP.

With other words, the VM may provide only one block of data per R-TP, while each R-TP may build for itself the redundant versions with the appropriate header. This involves that if the redundant packets are built by the R-TP, adding the CRC to the TB is a function of the R-TP.

Another possibility of functional partition is that the VM provides all the possible CRCs, one per re-transmission, but only one data packet. The R-TP does not compute the CRC by itself, but appends at each retransmission the suitable CRC.

In relation to ARQ: the UE sends a status report indicating the missing packets or fragments. This information arrives first at R-TP and may be forwarded to VM. The R-TP or the VM, based on the stored information, can build the TB for re-transmission. If the VM does this, the delay is higher.

C3, function of real-time delay and jitter measurements and also based on information regarding the VM loading and processing delay, can decide where the TBs for retransmission are formed: at VM or at R-TP.

From the above discussion results that a message should be transmitted by the VM or by the VM platform controller to the C3, indicating the delay and jitter between UE-specific entities (RLC, MAC, PDCP) and each relevant R-TP. For reducing the amount of information, in case that multiple UEs are handled in the same VM, the C3 should be provided by the VM Controller with the information, for each VM ID (VM Identifier), of the processed UEs. In such a case, the delay, jitter and load information can be provided per VM.

6 Conclusions

The capability of exposing the network state and user conditions across heterogeneous radio access networks in a unified and RAT-agnostic manner is essential for effective control and coordination of networking resources. Developing this capability is a challenging task for several reasons: 1) the type of parameters available may differ substantially between RATs; 2) the associated parameter definitions and ranges are in many cases vendor-specific, and 3) common abstractions used in multi-RAT settings require additional design considerations to ensure that the observed network state or performance is comparable between different RATs [94].

In WP3, we have addressed these challenges and explored novel combinations of existing metrics, supporting coordinated controller actions by higher-level abstractions in software-defined RANs, covering six objectives (Table 1-1 in section 1). In summary, we have:

- produced 4 contributions for joint prototyping WP6 and/or SDK code release (Sec. 1)
- defined the COHERENT information sharing model and the concept of network graphs (Sec. 2)
- identified and derived new expressions for 21 abstractions enabling resource coordination and control across heterogeneous RATs (Sec. 3);
- evaluated novel controller applications using proposed abstractions in 16 technical contributions on coordinated control and communication in 5G (Sec. 4)
- proposed an X2-like protocol for control and coordination of 3GPP RATs and WiFi (Sec. 5)

The COHERENT information sharing concept developed in WP2 and WP3 encompasses a decentralized, hierarchical, and recursive structure of abstracted information processed and aggregated at different levels of the architecture via controller agents, thereby supporting distributed and locally centralized control. Proposed abstractions listed in section 3 are RAT-agnostic, in the sense that control actions can be expressed at a high-level of abstraction and translated down to RAT-specific control actions.

Contributions in WP3 propose novel abstractions and controller applications at both RTC and C3 levels. Several abstractions used for network performance exposure are also generically applicable for various kinds of control operations (e.g. load balancing and SLA/SLO enforcement). In general, the abstractions are designed to be scalable and resource efficient, supporting distributed control operation as well as observability (local and global network views) by the use of local aggregation models and estimation. The general modelling approach for all contributions is based on composing low-level, basic metrics into composite metrics enabling new types and expressions of high-level abstractions generic to a wide range of RATs and controller applications. In section 3, we list common RAT-agnostic abstractions developed for various controller applications for resource coordination and QoS assurance, load balancing and D2D communication (Sec. 4).

The proposed abstractions and controller applications were evaluated mainly on LTE and WiFi networks. The generality of the approaches and extendibility to other types of RATs have in this deliverable mainly been qualitatively analyzed and assessed. Most contributions to 3GPP and activities in ETSI BRAN are based on D3.2 content, and in particular on sections 2.3, 5, and Annexes, which constitute a WP3 output to WP7. Next step involves further evaluation of the gain in using WP3 concepts by the integration of a subset of the proposed methods in WP6 [95], partially based on the implemented contributions to the COHERENT SDK developed in WP2 [96] and to algorithms in WP5 [7].

References

[1] COHERENT DoW

[2] COHERENT D2.1, “Use cases and architecture”, v1.0, 31/01/2016.

[3] COHERENT D2.2, “System Architecture and Abstractions for Mobile Networks”, v1.0, 20/07/2016.

[4] COHERENT D3.1, “First report on physical and MAC layer modelling and abstraction”, v1.0, 02/07/2016.

[5] COHERENT M3.2, “Modelling, abstraction and control of D2D communications done as inputs for WP2, WP4, WP5 and WP7”, v1.1, 25/10/2017.

[6] COHERENT D5.1, “Draft specification and implementation of the algorithm from programmable Radio Access Networks”, v1.0, 16/08/2017.

[7] COHERENT D5.2, “Final specification and implementation of the algorithm from programmable Radio Access Networks”, in preparation.

[8] COHERENT D6.1 “Draft report on technical validation” V1.0, 24.03.2016.

[9] A. Cipriano, “Modelling of D2D communications in LTE-Advanced Pro”, COHERENT confidential Technical Report, v1.0, 05/10/2017.

[10] 3GPP TS 23.501, “System Architecture for the 5G System; Stage 2”, Rel.15.

[11] 3GPP TS 23.203, “Policy and charging control architecture, Draft, Rel 14.

[12] 3GPP TS 36.300, “NR and NG-RAN Overall Description; Stage 2”, Rel. 15.

[13] 3GPP TS 36.214, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements", v14.1.0, February 2017.

[14] ETSI, “Network Functions Virtualisation (NFV)”, White Paper #3, at the “SDN & OpenFlow World Congress”, Dusseldorf, Germany, October 14-17, 2014.

[15] 3GPP TR 38.913 “Study on Scenarios and Requirements for Next Generation Access Technologies”; Rel. 14.

[16] S. Sesia, I. Toufik and M. Baker, LTE - The UMTS Long Term Evolution - From Theory to Practice, second edition, John Wiley & Sons, 2011.

[17] I. Latif, F. Kaltenberger, R. Knopp, J. Olmos, “Low Complexity Link Abstraction for MIMO Transmission in LTE/LTE-Advanced with IR-HARQ”, COST IC1004 TD(12)05060, Bristol, UK, 24-26 September, 2012.

[18] A. Cipriano et al., “Calibration issues of PHY layer abstractions for wireless broadband systems,” in Proc. IEEE VTC, Sep. 2008, pp. 1-5.

[19] LiangCheng Jiang, Fengyi Jiang, "A comprehensive and practical model for HARQ in LTE system", Wireless Communications and Networking Conference (WCNC) 2013 IEEE, pp. 901-905, 2013, ISSN 1525-3511.

[20] J.-F. Cheng, A. Nimbalker, Y. Blankenship, B. Classon, and T. K. Blankenship, “Analysis of Circular Buffer Rate Matching for LTE Turbo Code, IEEE 68th Vehicular Technology Conference (VTC 2008-Fall), Calgary, Canada, 21-24 Sept. 2008.

[21] A. Cipriano, and D. Panaitopol, “Performance Analysis of Sidelink Data Communications in Autonomous Mode”, submitted to WCNC 2018.

[22] A. Anttonen et al., “Finite-horizon prediction of energy depletions in off-grid wireless networks,” IEEE Trans. Veh. Tech., Aug. 2016.

[23] J. Andrews et al., "An overview of load balancing in hetnets: old myths and open problems," IEEE Wireless Commun., vol.21, no.2, pp.18-25, April 2014.

[24] Davit Harutyunyan, Supreeth Herle, Dimitri Marandin, George Agapiu‡, and Roberto Riggio, "Traffic–Aware User Association in Heterogeneous LTE/WiFi Radio Access Networks", in Proc. of IEEE NOMS 2018, Taipei, Taiwan.

[25] P. Kreuger, O. Görnerup, D. Gillblad, T. Lundborg, D. Corcoran and A. Ermedahl. (2015) “Autonomous Load Balancing of Heterogeneous Networks”, in IEEE 81st Vehicular Technology Conference (VTC Spring pp 1-5).

[26] P. Kreuger, R. Steinert, O. Görnerup, and D. Gillblad, “Distributed dynamic load balancing with applications in radio access networks,” International Journal of Network Management, Nov. 2017.

[27] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, ANSI/IEEE Std 802.11, LAN/MAN Standards Committee of the IEEE Computer Society Std., 2012

[28] R. Riggio, M. K. Marina, J. Schulz-Zander, S. Kuklinski, and T. Rasheed, “Programming abstractions for software-defined wireless networks,” IEEE Transactions on Network and Service Management, vol. 12, no. 2, pp. 146–162, 2015.

[29] L. Suresh, J. Schulz-Zander, R. Merz, A. Feldmann, and T. Vazao, “Towards Programmable Enterprise WLANS with Odin,” in Proc. of ACM HotNets, New York, 2012.

[30] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Amendment 1: Radio Resource Measurement of Wireless LANs, ANSI/IEEE Std 802.11k, LAN/MAN Standards Committee of the IEEE Computer Society Std., 2008

[31] B. Pfaff, J. Pettit, K. Amidon, M. Casado, T. Koponen, and S. Shenker, “Extending Networking into the Virtualization Layer,” in Proc. of ACM Workshop on Hot Topics in Networks, New York, 2009.

[32] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. F. Kaashoek, “The Click Modular Router,” ACM Trans. Comput. Syst., vol. 18, no. 3, pp. 263–297, 2000.

[33] D. Xia, J. Hart, and Q. Fu, “Evaluation of the Minstrel rate adaptation algorithm in IEEE 802.11g WLANs,” in Proc. of IEEE ICC, Budapest, 2013

[34] ISO/IEC and ITU-T, “High Efficiency Video Coding (HEVC). ITU-T Recommendation H.265 and ISO/IEC 23008-2 (version 4),” Dec. 2016.

[35] “FFmpeg project,” 2016. [Online]. Available: https://ffmpeg.org/.

[36] 3GPP TR 23.733, “Study on architecture enhancements to ProSe UE-to-Network Relay”, Draft, planned for Rel 15.

[37] 3GPP TS 36.211, “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation”, Rel 13.

[38] 3GPP TS 36.213, “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 14)”, v14.1.0, December 2016.

[39] 3GPP TS 36.331, “Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification”, Rel 13.

[40] D. Panaitopol, “Mobility study for high speed platform”, COHERENT confidential Technical Report, v1.0, 05/10/2017.

[41] Vlavianos, Angelos, et al. "Assessing link quality in IEEE 802.11 wireless networks: which is the right metric?." Personal, Indoor and Mobile Radio Communications, 2008. PIMRC 2008. IEEE 19th International Symposium on. IEEE, 2008

[42] NGMN Alliance, “NGMN 5G White Paper,” Mar. 2015.

[43] A. Ravanshid et al., “Multi-connectivity functional architectures in 5G,” 2016 IEEE International Conference on Communications Workshops (ICC), Kuala Lumpur, 2016, pp. 187-192.

[44] M. Giordani, M. Mezzavilla, S. Rangan and M. Zorzi, "Multi-connectivity in 5G mmWave cellular networks," 2016 Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net), Vilanova i la Geltru, 2016, pp. 1-7.

[45] M. Riza Akdeniz et al., “Millimeter Wave Channel Modeling and Cellular Capacity Evaluation,” IEEE JSAC, vol. 32, no. 6, Jun. 2014, pp. 1164-1179.

[46] H. Zhao et al., “28 GHz Millimeter Wave Cellular Communication Measurements for Reflection and Penetration Loss in and around Buildings in New York City,” Proc. IEEE ICC, Jun. 2013, pp. 5163-5167.

[47] T. Bai et al, “Coverage and Rate Analysis for Millimeter-wave Cellular Networks,” IEEE Trans. Wireless Commun., vol. 14, no. 2, Feb. 2015, pp. 1100-1114.

[48] H. Shokri-Ghadikolaei et al., “Millimeter Wave Cellular Networks: A MAC Layer Perspective,” IEEE Trans. Commun., vol. 63, no. 10, Oct. 2015, pp. 3437-3458.

[49] P. Mach, Z. Becvar, and T. Vanek, “In-band device-to-device communication in OFDMA cellular networks: A survey and challenges,” IEEE Commun. Surveys Tuts., vol. 17, no. 4, pp. 1885-1922, 4th Quart., 2015.

[50] G. Fodor, S. Roger, N. Rajatheva, S. B. Slimane, T. Svensson, P. Popovski, J. M. B. Da Silva and S. Ali, “An Overview of Device-to-Device Communications Technology Components in METIS,” IEEE Access, vol. 4, pp. 3288-3299, Jun. 2016.

[51] J. Deng, A. A. Dowhuszko, R. Freij, and O. Tirkkonen, “Relay selection and resource allocation for D2D-relaying under uplink cellular power control,” in Proc. IEEE Globecom Workshops (GC Wkshps), Dec. 2015.

[52] R. K. Ganti and M. Haenggi, “Spatial analysis of opportunistic downlink relaying in a two-hop cellular system,” IEEE Trans. on Commun. , vol. 60, pp. 1443-1450, May 2012.

[53] A. Moller and U. T. Jonsson, “Stability of rate and power control algorithms in wireless cellular networks,” in Proc. IEEE Decision and Control and European Control Conference, pp. 4535-4541, Dec. 2011.

[54] J.-M. Kelif, M. Coupechoux, and P. Godlewski, “A Fluid Model for Performance Analysis in Cellular Networks,” EURASIP Journal on Wireless Commun. and Netw., Aug. 2010.

[55] H. ElSawy, A. Sultan-Salem, M. S. Alouini and M. Z. Win, “Modeling and Analysis of Cellular Networks Using Stochastic Geometry: A Tutorial,” IEEE Commun. Surveys Tuts., vol. 19, no. 1, pp. 167-203, 1st Quart. 2017.

[56] J. Deng, O. Tirkkonen, T. Chen, “D2D Relay Management in Multi-cell Networks”, IEEE ICC 2017 Wireless Communications Symposium.

[57] Y. Li and A. Ephremides, “A joint scheduling, power control, and routing algorithm for ad hoc wireless networks,” Ad Hoc Networks, vol. 5, no. 7, pp. 959 – 973, 2007

[58] D. Peethala, N. Zarifeh and T. Kaiser, “Probability of Coverage Based Analysis of Distributed Antenna System and its Implementation on LTE Based Real-Time-Testbed”, 2017, The 9th International Congress on Ultra Modern Telecommunications and Control Systems ICUMT, Munich, 2017

[59] 5G PPP Architecture Working Group, “View on 5G Architecture”, Jul. 2016. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-July-2016.pdf

[60] NGMN, “5G initiative white paper,” NGMN Alliance, Tech. Rep., Feb. 2015. [Online]. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf

[61] N. Zarifeh, A. Kabbani, M. El-Absi, T. Kreul and T. Kaiser, "Massive MIMO exploitation to reduce HARQ delay in wireless communication system," 2016 IEEE Middle East Conference on Antennas and Propagation (MECAP), Beirut, 2016, pp. 1-5.

[62] 3GPP, “Feasibility study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN),” 3rd Generation Partnership Project (3GPP), Tech. Rep. 25.912, Sep. 2012.

[63] M. T. Kawser, N. B. Hamid, M. N. Hasan, M. S. Alam, and M. M. Rahman, “Limiting HARQ Retransmissions in Downlink for Poor Radio Link in LTE,” International Journal of Information and Electronics Engineering, vol. 2, no. 5, pp. 707–709, 2012.

[64] B. Bossy, P. Kryszkiewicz, H. Bogucka "Optimization of Energy Efficiency in the Downlink LTE Transmission" IEEE International Conference on Communications 2017

[65] Z. Hanzaz and H. D. Schotten, "Performance evaluation of Link to system interface for Long Term Evolution system," 2011 7th International Wireless Communications and Mobile Computing Conference, Istanbul, 2011, pp. 2168-2173.

[66] P. Kryszkiewicz, A. Kliks and H. Bogucka, "Small-Scale Spectrum Aggregation and Sharing," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 10, pp. 2630-2641, Oct. 2016.

[67] R. Kwan, C. Leung and J. Zhang, "Proportional Fair Multiuser Scheduling in LTE," in IEEE Signal Processing Letters, vol. 16, no. 6, pp. 461-464, June 2009.

[68] Holland, Gavin, Nitin Vaidya, and Paramvir Bahl. "A rate-adaptive MAC protocol for multi-hop wireless networks." Proceedings of the 7th annual international conference on Mobile computing and networking. ACM, 2001.

[69] Kawser, Mohammad T., et al. "Downlink SNR to CQI Mapping for Different MultipleAntenna Techniques in LTE." International Journal of Information and Electronics Engineering 2.5 (2012): 757.

[70] Sesia, Stefania, Matthew Baker, and Issam Toufik. LTE-the UMTS long term evolution: from theory to practice. John Wiley & Sons, 2011.

[71] Moravapalle, Uma Parthavi, et al. "Pulsar: improving throughput estimation in enterprise LTE small cells." Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies. ACM, 2015.

[72] NS3 https://www.nsnam.org/

[73] D. W. Griffith, F. J. Cintron, and R. A. Rouil, “Physical Sidelink Control Channel (PSCCH) in Mode 2: Performance Analysis”, in 2017 IEEE International Conference on Communications (ICC 2017), May 2017.

[74] R. Heath, S. Peters, Y. Wang, and J. Zhang, ”A current perspective on distributed antenna systems for the downlink of cellular systems,” IEEE Commun. Mag., vol.51, no.4, pp.161-167, April 2013

[75] L. Dai, S. Zhou, and Y. Yao, ”Capacity analysis in CDMA distributed antenna systems,” IEEE Trans. Wireless Commun., vol.4, no.6, pp.26132620, Nov. 2005

[76] H. Osman, H. Zhu, D. Toumpakaris, and J. Wang, ”Achievable rate evaluation of in-building distributed antenna systems,” IEEE Trans. On Wireless Communications vol. 12, no. 7, pp. 35103521, Nov. 2013

[77] X.-H. You, D.-M. Wang, B. Sheng, X.-Q. Gao, X.-S. Zhao, and M. Chen, Cooperative distributed antenna systems for mobile communications, IEEE Wireless Communications pp. 3543, Jun. 2010

[78] W. Choi, and J. G. Andrews, "Downlink performance and capacity of distributed antenna systems in a multicell environment," IEEE Trans. Wireless Commun., vol.6, no.1, pp.69-73, Jan. 2007

[79] R. Bosisio, and U. Spagnolini, "Interference Coordination Vs. Interference Randomization in Multicell 3GPP LTE System," IEEE Wireless Communications and Networking Conference, 2008. WCNC 2008., vol., no., pp.824-829, March 31 2008-April 3 2008

[80] Z. Liu, and L. Dai, "A Comparative Study of Downlink MIMO Cellular Networks With Co-Located and Distributed Base-Station Antennas," IEEE Trans. Wireless Commun., vol.13, no.11, pp.6259-6274, Nov. 2014

[81] K. T. Truong, and R. W. Heath, "The viability of distributed antennas for massive MIMO systems," Asilomar Conference on Signals, Systems and Computers, 2013, vol., no., pp.1318-1323, 3-6 Nov. 2013

[82] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO: Uniformly great service for everyone," IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2015, vol., no., pp.201-205, June 28 2015 - July 1 2015

[83] A. A. M. Saleh, A. J. Rustako, and R. Roman, "Distributed Antennas for Indoor Radio Communications," IEEE Trans. Commun.", vol.35, no.12, pp.1245-1251, December 1987.

[84] Y. D. Beyene, R. Jantti, and K. Ruttik, "Cloud-RAN Architecture for Indoor DAS," IEEE Access, vol.2, no., pp.1205-1212, 2014

[85] K. Heejin, S. Lee, and I. Lee, "Sum Rate based Transmission Selection Schemes in Distributed Antenna Systems," Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th, Dresden, 2013 pp. 1-5

[86] K. Govindan, K. Zeng and P. Mohapatra, "Probability Density of the Received Power in Mobile Networks," IEEE Transactions on Wireless Communications , vol. 10, no. 11, pp. 3613-3619, November 2011.

[87] 3GPP TS 36.423 v.14.0.0; Evolved Universal Terrestrial Radio Access (E-UTRA); Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 application protocol (X2AP) (Release 14).

[88] J. Deng, O. Tirkkonen, R. Freij-Hollanti, T. Chen and N. Nikaein, "Resource Allocation and Interference Management for Opportunistic Relaying in Integrated mmWave/sub-6 GHz 5G Networks," in *IEEE Communications Magazine*, vol. 55, no. 6, pp. 94-101, 2017.

[89] N. Y. Ermolova and O. Tirkkonen, "Interference Analysis in Wireless Networks Operating over Arbitrary Fading Channels with Heterogeneous Poisson Fields of Transmitters and Interferers," in *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1388-1392, Sept. 2017.

[90] Pawel Kryszkiewicz „Multi-RAT Scheduling for Heterogeneous Networks" IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 8-13 October 2017, Montreal, Canada

[91] Rajesh Mahindra, Hari Viswanathan, Karthik Sundaresan, Mustafa Y. Arslan, and Sampath Rangarajan. 2014. A practical traffic management system for integrated LTE-WiFi networks. In *Proceedings of the 20th annual international conference on Mobile computing and networking* (MobiCom '14). ACM, New York, NY, USA, 189-200. DOI=http://dx.doi.org/10.1145/2639108.2639120.

[92] S. Lembo, J. Deng, R. Freij-Hollanti, O. Tirkkonen, and T. Chen. “Hierarchical Network Abstraction for HetNet Coordination”. In IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, Oct. 2017.

[93] A. Rao, R. Steinert. “AQuaMet: Network Interface Selection using Unified Probabilistic Network Abstractions”. Submitted for review in IEEE International Conference on Communications (ICC), 2018.

[94] Git repository for WiFi and LTE NIF code, “https://github.com/nigsics/aquamet”.

[95] R. Steinert, N. Zarifeh, C.Y. Chang, A. Anttonen, A. Cipriano, D. Panaitopol, "Toward unifying abstractions for heterogeneous radio infrastructures," European Conference on Networks and Communications (EuCNC’2017), Oulu, Finland, June 12-15, 2017.

[96] COHERENT D6.2, “Final report on technical validation”.

[97] COHERENT D2.4, “Final release of the SDK”.

Annex A. Public Material

Abbreviations

BSS Basic Service Set (as used in [1])

SME Station Management Entity

MLME MAC subLayer Management Entity

A.1 Control framework for heterogeneous IEEE 802.11 – 3GPP LAA network

A.1.1 New Information Elements for general reporting

This section describes the new Information Elements; most of the messages are new in 3GPP.

Each node sends messages to C3 indicating at least one of the following information:

a. *Received power histogram in energy detection*; for both 3GPP and WiFi: If it is not possible to decode the transmission node ID, the listening node will just sample the received power and memorize it together with a time stamp, creating a histogram.

b. *Release support capability*; for both 3GPP and WiFi: A code for the supported technologies (e.g. LTE release number and type of operation, WiFi indicative of technology or 802.11 indicative of technology, for example the amendment numbers, like 802.11ac);

c. *Used channel*: for both 3GPP and WiFi: A code for the used channel;

d. *Power usage indication*: for both 3GPP and WiFi: Maximum and average transmission power per node or per channel;

e. *DL user specific MCS log file*: for both 3GPP and WiFi: A log file containing the MCS of past transmissions to specific UEs and the transmission start time and duration;

f. *UL user specific MCS log file: f*or both 3GPP and WiFi: At least one measurement or a log file containing the MCS of past receptions from specific UEs and the start time and duration of received data;

g. *UL/DL CSI/CQI/RS-SINR log file or statistical information*: for 3GPP: At least one measurement or medium or maximum or minimum values or a log file containing the CSI or CQI or RS-SINR of the past receptions from specific nodes and the start time and duration of received data;

h. *Interference power log file or statistical information in active or idle state:* for 3GPP: At least one measurement or medium or maximum or minimum values or a log file containing the interference power measured on zero-power CSI-RS reference signals of past receptions or past idle state of specific nodes and the start time and duration of the measurement;

i. *RSSI log file or statistical information in active or idle state:* For both 3GPP and WiFi: At least one measurement or medium or maximum or minimum values or a log file containing the power measured as RSSI of past receptions or past idle state of specific nodes on the operating channel(s) and the start time and duration of the measurement;

j. *RSSI log file or statistical information measured on non-operating channels*: for both 3GPP and WiFi: At least one measurement or medium or maximum or minimum values or a log file containing the power measured as RSSI of past idle state of specific nodes on at least one non-operating channel and the start time and duration of the measurement;

k. *Percentage of channel usage relative to the energy detection threshold*: for 3GPP: the percentage of time in which each monitored channel was busy (relative to energy detection threshold or carrier sense threshold) or it was free;

l. *Average and peak user throughput*: for both 3GPP and WiFi average throughput and peak throughput per user or total;

m. *Delay due to congestion*; for both 3GPP and WiFi: the delay due to congestion, i.e. due to retransmissions;

n. *Number or percentage of discarded packets*: for 3GPP

o. *Received power from other nodes as log file or statistical information:* for both 3GPP and WiFi: received power from other nodes either in a statistical form (average, point on CDF) or as a list of samples.

p. *Received power and the neighbor transmitting node ID*: for both 3GPP and WiFi: when possible, the received power and the neighbor transmitting node ID.

A.1.2 Reports for supporting QoS enforcement per flow or per radio bearer

Part of 3GPP QoS enforcement, we envisage the reports from the UE/BS to C3 regarding:

a. *Actual Average and maximum Flow Bit Rate* or a CDF representative value (e.g. 10%, 90%) for the Flow Bit Rate in the measurement interval

b. *Actual CW Priority level*, reflected by a priority index or the average and the minimum and maximum CW (Contention Window) size back-off rules

c. *Average and maximum flow Packet Delay* or a CDF representative value (e.g. 10%, 90%) for the Average and maximum or a CDF representative value for the Packet Error rate.

d. *Average and maximum per-transmission flow medium occupancy* or a CDF representative value for the per-transmission flow medium occupancy

e. *Application type*, possibly represented by a flow ID or a PDU session ID, and if available an estimation of the duration of the active state is reported to the C3, for improving its decisions.

f. *Actual QoS parameters per Service Data Flow* which may include:

- Maximum Flow Bit Rate
- Guaranteed Flow Bit Rate: the bitrate (Minimum or Guaranteed bitrate per flow) that is required for the service to be delivered with sufficient QoE
- Priority level
- Packet Delay Budget
- Packet Error rate
- Admission control.

The infrastructure nodes may relay or tunnel to C3 the reports of the stations of UEs.

Such reports may include per traffic type or per a specific flow:

a. The power levels for transmission and the time-stamp of the transmission start and end
b. The power levels measured during reception, including a time-stamp and sampled at regular intervals and presented as a file log or as representative points on a CDF
c. The SINR during reception, including a time-stamp and sampled at regular intervals and presented as a file log or as representative points on a CDF
d. The power levels measured during idle intervals, including a time-stamp and sampled at regular intervals and presented as a file log or as representative points on a CDF
e. MCS (modulation and coding state) used for transmission
f. The used Listen Before Talk (LBT) thresholds
g. The used priority level reflected by the maximum CW
h. Defer duration.

A.1.3 Registration to Central Coordinator (C3) and initial operation

In the centralized approach, before starting the operation, each infrastructure node (AP, base station) must listen to the environment and also register with the C3 by the set-up of a logical interface for communication.

By listening to the environment, it will be possible for a node to detect surrounding nodes and the RSSI (received signal strength indication) power. Depending on the supported technologies, for some of the nodes it will be possible to detect the surrounding node ID or Cell ID as well.

The information provided at interface set-up can include one or more information elements (IEs) containing information about:

a. identifier of the node
b. node network address, for example the IP address
c. location name or GPS position of the node
d. supported standards or technologies
e. supported frequencies of operation
f. type of LBT used (energy detection, which can be with fixed threshold or variable threshold, carrier sense, which can be with fixed threshold or variable threshold
g. capability of LBT per carrier in carrier aggregation
h. capability to modify the LBT level (energy detection level or carrier sense level)
i. medium occupancy limit
j. capability of operation related to channel width
k. capability of beamforming
l. maximum transmitted power
m. antenna gain
n. type of implementation: virtualized with multiple remote radio units, their ID, their location coordinates
o. operator name/ID
p. support of infrastructure sharing between operators
q. supported network slices if any
r. neighbor nodes ID
s. received RSSI from a node.

A.1.4 Operational performance

The C3 will assess based on a trigger event or periodically the wireless network operation performance based on the reports received from the different nodes.

A.1.5 Interference coupling

If the transmitting node provides to C3, through a message, the time stamp of the transmission start and stop, and the transmission power, the C3 will request from the other nodes the reports regarding the received power during these intervals. The reports may provide the sampled power

levels or a statistical representation including the average power and the power at some specific points on the CDF, for example 90%.

Based on this information C3 may create a better representation on the interference created by a node to other nodes from the same or different technologies and calculate the coupling loss between the transmitting and the other nodes.

The coupling loss, defined as the ratio between the transmitted power and the received power, will allow for example to limit the transmission power for increasing the frequency reuse factor or to request the use of a low LBT detection threshold for increasing the coverage of other node or for increasing the SINR for delivery of higher user throughput or higher probability of successful reception.

A.1.6 Dependency on the traffic type

Each user application may have specific requirements, for example high data rates for streaming video, low latency for V2X, high reliability for IoT.

IEEE 802.11 uses Access Categories (AC) for establishing the minimum and maximum size of the contention window (CW) and the exponential backoff of the CW for accessing the medium. The AC is based on the QoS, derived from DiffServ DSCP.

In the existing cellular networks, such as LTE, the QoS is marked by QCI; a different bearer is constructed based on QCI and the QCI values are used for establishing priorities.

In 3GPP, the traffic generated by an application or its agents will be categorized in both downlink and up-link into flows using TFT (Traffic Flow Template) packet filters (for example source IP address, port number, destination ip address, flow label, eventually IP protocol type) and each such flow can have an identifier to be carrier by each transported packet.

In the next generation 5G networks the Service Level Agreement establishes policy rules for the transport of a traffic flow, identified by a flow ID. The policy rules may alter the priorities carried by the native DSCP marking or by the QoS requirements of the application.

In order to provide the delivery of traffic such to match both the policy rules and the native QoS requirements, it is needed that C3 will be aware of the flow ID, policy rules and the native QoS for a specific traffic. A policy rule can be the values for the maximum throughput or the maximum delay or the reliability (percentage of discarded packets), established by the network based on QoS requirements of the application.

A concept new at this time in 3GPP is the definition of PDU sessions, where each PDU session is mapped to a separate transport network bearer in order to separate them even if the contained packets have an overlapping IP address range. Also, the UE must be able to determine which IP packet belongs to which PDU session in order to route packets correctly. There may be several PDU sessions per UE.

The traffic belonging to a PDU session may be served by several technologies, for example WiFi and LTE or other cellular technology.

Based on our approach, the actual QoS parameters achieved by each used technology in transporting the data belonging to a PDU session are reported to the C3 via a message.

The access stratum maps the bearers or flows to the appropriate DRB (Data Radio Bearer), each DRB having a dedicated queue, which is of direct relevance to the QoS over radio.

A PDU session could be mapped to a network slice, including a number of policy rules for resource allocation.

In our approach the policy rules are translated into radio parameters for operation in a shared band.

While the translation of the policy rules can be done locally in a distributed mode, a C3 entity having the full view of the network can make this translation in an optimal mode.

In particular, a node may request C3 through a message a change of LBT thresholds or transmitted power thresholds or delay or retransmission number or a repetition coding for itself or for the other nodes in the network, such to obtain the desired level of MCS (Modulation and Coding state) for a specific running application or a specific PDU session or a specific flow used in a PDU session. The request may include the duration of the change.

A1.3.5 C3 actions for QoS enforcement

The main medium access parameters are: energy or carrier sense power or energy threshold, maximum medium occupancy duration, a rule specifying the maximum contention window and the transitions in case of exponential back-off, value of the defer time.

Based on these reports and of the requirements resulting from the QoS related to PDU sessions, flows, DSCP marking, the C3 may change the allocated frequency channels to each node, transmission power threshold, and/or LBT thresholds, priority level including the CW rules and the defer duration.

A.1.3.6 Virtual LBT

C3 may ensure that a channel is free by sending a message to the relevant nodes in which it requests that a specific channel or channels or part of a channel will not be used for a specified period of time, such to allow transmissions of nodes having delay or throughput or reliable delivery problems**. We call this form of resource reservation Virtual LBT.**

The virtual LBT duration can be negotiable, for example based on technical arguments or on pricing.

Virtual LBT can be used for enabling higher SINR for selected IoT traffic, or PDU session or flow or bearer.

In this approach C3 can make available a channel for the operation of nodes using a technology by asking one or more nodes using another technology to reduce the occupancy of the channel by a given amount.

The occupancy of a channel can be expressed as the percentage of time in which a base station communicates with the associated UEs or as the percentage of time in which an AP communicates with the associated stations or as percentage of time in which stations communicate between them or as percentage of time in which UEs communicate between them.

A.1.3.7 Carrier aggregation and LBT thresholds

Both 802.11 and LTE use carrier aggregation; in 802.11 the LBT is used on each carrier separately.

C3 may set different LBT thresholds on each aggregated carrier and set the primary control channel for each UE/STA on a frequency channel selected by it. The effect of this setting is that in presence of interference above threshold only a part of the channels will be used.

The LBT threshold selection for a component carrier in Carrier Aggregation may be dependent on the application type using that carrier, for example a broadcast application may require high coverage hence a low LBT threshold from the neighbor transmitters.

The allocation of the frequency channels to be used for transmission by different nodes can be dependent on the data rate required by the application.

If low latency and low traffic throughput is requested at one AP/BS to communicate at relatively short distance but a high latency can be accepted by the BS/APs in vicinity, the best policy is to use a channel with high LBT threshold for the application requiring low latency and high LBT and multiple channels for the neighboring nodes. The exact allocation depends also on the scheduling policy, for example LTE can use per-carrier scheduling as a mode of carrier aggregation.

MCS selection

Given that C3 has a better view of the future interference, as result of the information received from the nodes and also based on its commands setting the power and LBT levels, and eventually the time-frequency resources reserved by the 802.11 virtual carrier sense, C3 can give a more accurate prevision of the MCS to be used by the nodes in their communication with other nodes.

The MCS selection can be specific for the application, such that an application requiring low delay can be allowed by C3 a higher SINR for increasing the chances of successful reception.

Communication between C3 and Application servers

An application server can provide C3 the possibility to communicate, through a protocol, and obtain from the application server the information regarding the requirements of the traffic at a more detailed level than provided by the packet marking by DSCP. For example, it can make a differentiation between low delay and high reliability.

Based on both DSCP and this information transmitted by the C3 through a message to UE/STA, the UE/STA can classify certain traffic as belonging to a specific network slice. **The specific network slice identifier can be used for setting threshold values for the LBT threshold and the transmitted power.**

IoT traffic

The IoT packets may be characterized by shortness and lack of sensitivity to delay, while their loss shall be very small; this allows a long TTL (time-to-live) in the network. We understand that this kind of behavior is not supported by the current IETF standards/RFCs and by the 3GPP QCI table in 3GPP TS 23.203.

The 3GPP QCI table should be improved by differentiating between priority, which also implies fast delivery, and the discard rate, i.e. new fields need to be introduced for reflecting only high reliability without time constraints.

This can be achieved by adding a QCI value for which, for example, either a very long TTL (Time To Live) or a condition for low discard time of the packet from the transmission queue is explicitly mentioned. In Table A.1 shows different fields for delay and loss.

Table A.1 QCI with differentiation between delay and loss

QCI	Resource Type	Priority	Packet Delay Budget	Packet Error Loss	Example Services
t.b.d.	Non-GBR	None	>300ms	10-6 or lower	High reliability IoT

Radar detection

Radar detection is required in many shared bands.

Given the specific radar pulse waveform, the detection is done during receive periods, as long as the energy of the radar signal is high enough relative to the energy of the background signals which interfere with the radar signals. Each AP/BS/STA/UE shall be aware of its performance in radar detection.

In case that a device becomes aware that the interference power and the interference duration make the radar detection impossible with the required precision, it shall announce this to C3 through a message; if the C3 can rely on another device in the area which can detect the radar, will send a message to the reporting device allowing its operation; otherwise the C3 shall take measures, by sending messages requesting the channel change, to move the interfering or the interfered device to another channel.

C3 and common protocol

To allow the adaptability of parameters for accessing a common medium by entities belonging to different technologies, it is used at least a common C3 and eventually a common protocol to be implemented by different wireless node belonging to different technologies sharing the medium. The common C3 shall optimize the access to the medium such that each node will have maximum performance in relation with the performance metrics established by an Operator or resulting from traffic and application or PDU session QoS requirements.

Alternatively, the Information Elements may be common to different technologies, but the communication protocol could be specific to the standardization group, i.e. SNMP for 802.11 and X2-like for 3GPP. We note that X2–like offers more diverse possibilities as compared with SNTP.

A.2 Control Framework of the NG User Plane

A.2.1 HO Option 1

Considering Figure A-1, while the RLC in DU uses TM or UM, the SN (Sequence Number) in the PDCP header for a PDU delivered by RLC to MAC layer (named SNt) is used to mark the last relevant PDU sent to UE, with the assumption that the UE PHY will not disconnect from the source DU until the last PHY SDU was sent successfully or the H-ARQ process expired. This SNt is used for identifying the SDU of the RLC layer, i.e. the PDU of the PDCP layer. “Relevant PDU” means a RLC PDU which corresponds to the specific PDU session, Slice identifier, QoS Flow identifier for which the HO is done.

When the RLC in DU uses AM, the SN in the PDCP header for a PDU delivered by RLC to MAC layer (named SNt) is used to mark the last PDU acknowledged by the RLC layer in UE and reported to the RLC layer in DU. This SN is used for identifying the SDU of the RLC layer.

The RLC layer sends the scalar representing SNt or (SNt +1) to the C3/CU' relevant controller / coordinator through a signaling message over the CU-DU interface.

All the SDUs having the PDCP SN higher than SNt will be forwarded by the PDCP layer or by the CU to the destination DU. Option 1 procedure is mostly suitable for low delay, low loss backhauls working in non-congestion state.

A.2.2 HO Option 2

In this option the delay or packet loss in the CU-DU link is high, while the delay between the source DU and the destination DU is lower and also the packet loss is lower as compared with CU-DU link.

The TLC/RLC shall buffer the SDUs for a time equal at least with the maximum backhaul delay. Based on the SNt obtained as in Option 1, the RLC layer forwards the SDUs having the SN of the PDCP layer higher than SNt directly to the destination DU and announces through a signaling message over the CU-DU interface the lowest and the highest SN of the forwarded SDUs. The message can be sent to the source PDCP entity and/or to the C3.

A.2.3 HO Option 3

In this option the delay or packet loss in the CU-DU link is high, while the delay between the source DU and the destination DU is lower and also the packet loss is lower as compared with CU-DU link.

The TLC shall buffer the SDUs for a time equal at least with the maximum backhaul delay.

The RLC forwards the SNt obtained as in option 1 to the TLC-B in the DU. Based on a configuration transmitted by a signaling control message, TLC-B or RLC in DU can either forward the PDCP packets as in Option 2 or use the PDCP SNt of the last successful reception by TLC-B, with the assumption that all the received relevant packets will be delivered to UE before the PHY HO.

The TLC-B forwards the SDUs having the SN of the PDCP layer higher than SNt directly to the destination DU and announces through a signaling message over the CU-DU the lowest and the highest SN of the forwarded SDUs. The message can be sent to the source PDCP entity and/or to the C3.

A.2.4 Short IP packets case

As shown in the presentation SITAPT, "Study of internet traffic to analyze and predict traffic", by Amit Arora, while the TCP traffic represents aprox. 80% of the packets, the "Packet size distribution is almost entirely bimodal, with only the 1 to 100 bytes range and the 1400 to 1500 bytes range showing packet percentages of any significance".

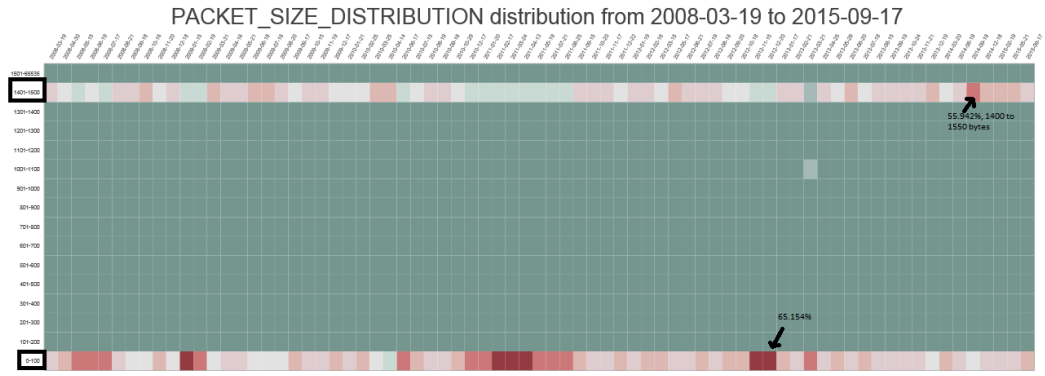


Figure A-1 Packet size distribution

When computing the approx. overhead of the short packets (100 bytes) relative to long packets (1400 bytes) for an equal percentage of each type of size, it results that the percentage of their overhead is 7.14%. Short packages may belong to:

- TCP ACK
- Voice
- Short messages
- IoT (Internet of Things)
- Cellular V2X (in future, cellular Vehicle to Everything).

The fact is that many of the short packets are discarded, impeding on the throughput and functionality of the application. On the other hand too many short packets may drop the system capacity, given the limitation of the computing power for processing the headers.

A.2.5 RLC and PDCP layers

A.2.5.1 RLC Layer

RLC operation as used in LTE is described in 3GPP TS 36.322 v13.2.0 (2016-06) for BS – UE (User Equipment) communication over the air and shortly described below.

Functions of the RLC sub layer are performed by RLC entities. For a RLC entity configured at the base station, there is a peer RLC entity configured at the UE and vice versa.

This standard defines three operational modes: Transparent Mode (TM), Unacknowledged Mode (UM) or Acknowledged Mode (AM). Based on this standard, an RLC entity can operate only in one of the 3 modes for all the types of handled traffic and a set of logical channel, such as control channels and traffic channels are assigned for each entity.

The ARQ is used only in the Acknowledged mode for DL/UL control channel or DL/UL traffic channel.

In the transparent mode the transmitter forms TMD (TM Data) PDUs from RLC SDUs, but will not segment nor concatenate the RLC SDUs and no RLC headers will be included in the TMD PDUs. The receiver delivers the TMD PDUs (which are just RLC SDUs) to the upper layer.

In the UM a transmitting UM RLC entity forms UMD PDUs from RLC SDUs, and shall segment and/or concatenate the RLC SDUs so that the UMD PDUs fit within the total size of RLC PDU(s) indicated by the lower layer at the particular transmission opportunity notified by lower layer and also include relevant RLC headers in the UMD PDU.

The receiving UM RLC entity receives UMD PDUs and detects whether or not the UMD PDUs have been received in duplication and discards duplicated UMD PDUs, reorders the UMD PDUs if they are received out of sequence, detects the loss of UMD PDUs at lower layers, reassemble RLC SDUs from the reordered UMD PDUs and delivers the RLC SDUs to upper layer in ascending order of the RLC SN (Sequence Number), discards the received UMD PDUs that cannot be re-assembled into a RLC SDU.

The transmitting side of an AM RLC entity forms AMD PDUs from RLC SDUs, and segments and/or concatenates the RLC SDUs so that the AMD PDUs fit within the total size of RLC PDU(s) indicated by lower layer at the particular transmission opportunity notified by lower layer.

Only in AM the transmitting side of an RLC entity supports retransmission of RLC data PDUs (ARQ).

If the RLC data PDU to be retransmitted does not fit within the total size of RLC PDU(s) indicated by lower layer at the particular transmission opportunity notified by lower layer, the AM RLC entity can re-segment the RLC data PDU into AMD PDU segments, the number of re-segmentation being not limited.

When the transmitting side of an AM RLC entity forms AMD PDUs from RLC SDUs received from upper layer or AMD PDU segments from RLC data PDUs to be retransmitted, it shall include relevant RLC headers in the RLC data PDU.

When the receiving side of an AM RLC entity receives RLC data PDUs, detects whether or not the RLC data PDUs have been received in duplication, and discard duplicated RLC data PDUs, reorders the RLC data PDUs if they are received out of sequence, detects the loss of RLC data PDUs at lower layers and requests retransmissions to its peer AM RLC entity, reassemble RLC SDUs from the reordered RLC data PDUs and deliver the RLC SDUs to upper layer in sequence.

It should be noted that in practice RLC has no memory, such that the retransmission is done by the PDCP layer, given that is easy to identify the PDCP SN corresponding to a RLC SN. Due to this, in CU-DU split the retransmission for a DL PDU will be done by the CU.

A.2.5.2 PDCP layer

As shown in Figure A-2, the higher layer for the TLC entity located on each side of CU-DU interface is PDCP; the higher layer for the RLC entity located in DU is also PDCP, as TLC is at the same level with RLC.

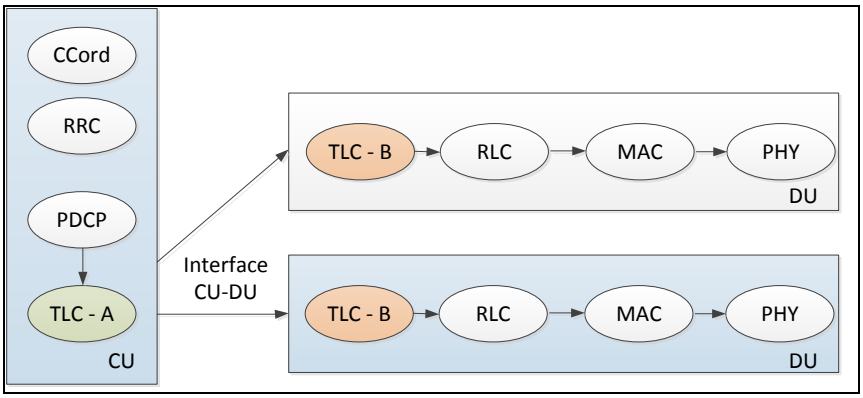


Figure A-2 TLC and high-layer split

The PDCP layer executes the following functions relevant for this approach: transfer of data (user plane or control plane), maintenance of PDCP SNs, in-sequence delivery of upper layer PDUs at re-establishment of lower layers; ROHC; duplicate elimination of lower layer SDUs at re-establishment of lower layers for radio bearers mapped on RLC AM; timer based discard; duplicate discarding; for split and LWA bearers, routing and reordering. PDCP uses the services provided by the RLC sublayer.

PDCP is used for SRBs, DRBs, and SLRBs mapped on DCCH, DTCH, and STCH type of logical channels. PDCP is not used for any other type of logical channels.

A.2.6 Details on TLC configuration

The configuration of the TLC instance which is not collocated with the configuring entity will be done by signaling messages belonging to the Control Plane and generated by the RRC entity or C3 indicating the TLC instance identifier and configuring:

1. The mode of processing, i.e. Transparent mode (TM), Un-Acknowledged mode (UM) or Acknowledged mode (AM).
2. The sequence number size
3. The number of retransmissions or the time-budget for retransmissions
4. The time-out to wait for an ACK
5. Policy rules in case when the number of lost packets or packet error rate or bit error rate or average delay or maximum delay is higher than a threshold.

An example of policy rule can be, for example, switching from AM to NAM in case of delay higher than a threshold and the PER is lower than a specific PER value.

An AM RLC entity sends STATUS PDUs to its peer AM RLC entity in order to provide positive and/or negative acknowledgements of RLC PDUs (or portions of them).

Alternatively, the signaling messages will assign a TLC/RLC instance to a specific TLC/RLC entity operating in a given mode.

A.3 Information Elements for F1-C Interface

This section is the continuation of section 5.2 in the main document, “Control Interface API for F1-C interface“.

A.3.1 Information Elements for Reports over F1-C

A.3.1.1 Time, frequency and space resources per cell in downlink

This IE provides an indication on the total number of time-frequency resources per cell and T-RP.

It is understood that a cell can be formed by multiple T-RPs.

The information is provided in a defined time interval, as the slot allocation for UL and DL transmission can change.

It is assumed that the C3 was configured by OEM with the number of RB per cell, such that this information is not retransmitted.

It is assumed that C3 has provided the maximum number of DL slots to be used by T-RP and per duplex frequency direction; however the T-RP may not intend to use all the slots.

The reports will use the statistical format as defined by C3 in the C3-INNITIATED ACTION REQUEST message.

Table A.2 Time, frequency and space resources per cell in downlink

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
Cell General Identifier in NR	M		Integer (1…4000000)	NCGI	YES	Reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
R-TP Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Start SFN	M		INTEGER (0..1023, …)	SFN of the radio frame containing the first sub-frame when the Time, frequency, and space resources per cell are valid.	–	–
Start Subframe Number	M		INTEGER (0..9, …)	Sub-frame number, within the radio frame indicated by the Start SFN IE.	–	–
Subframe Assignment to R-TP in DL (Time resources for DL)	M		BITSTRING (SIZE(32))	Each position in the bitmap indicates a slot, “’1” indicates that the slot is used for DL transmission. First bit indicates the first slot, 2nd bit the second slot and so on. The report uses the report format in the initial C3 RESOURCE STATUS REQUEST message.	YES	ignore
Pattern repetition interval	O		INTEGER	Indicated as number of subframes	-	-
Number Of Cell-specific Antenna Ports	M		ENUMERATED (1, 2, 4, …)	P (number of antenna ports for cell-specific reference signals) t.b.d. in TS 38.211	–	–
P_B	M		INTEGER (0..3, …)	t.b.d in TS 38.213	–	–

A.3.1.2 Summary of the resource availability per cell

The LTE Release 13 defines the use of two power thresholds for reporting the power reservation policy. As result, the summary of the available resources per cell should be presented as:

- Number of available non-restricted PRB resources in all the assigned sub-frames, i.e. the resources which are not restricted from p.o.v of transmitted power;
- Number of available fully restricted PRB resources in all the assigned sub-frames, i.e. resources with transmitted power below the lowest threshold;
- Number of available medium restricted PRB resources in all the assigned sub-frames, i.e. resources below a second threshold higher than the first power threshold.

The reports will use the statistical format as defined by C3 in the C3-INNITIATED ACTION REQUEST message.

Table A.3 Summary of the resource availability per cell

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
Cell General Identifier in NR	M		Integer (1…4000000)	NCGI	YES	Reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
R-TP Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Start SFN	M		INTEGER (0..1023, …)	SFN of the radio frame containing the first subframe when the Time, frequency, and space resources per cell are valid.	–	–
Start Subframe Number	M		INTEGER (0..9, …)	Subframe number, within the radio frame indicated by the Start SFN IE.	–	–
End Subframe Number	M		INTEGER (0..9, …)	Last considered sub-frame within the radio frame	YES	ignore
Non-restricted PRB resources	O		INTEGER	Total number of available PRBs which are not restricted by transmitted power. The statistical format is as defined by C3 in the C3-INNITIATED ACTION REQUEST message.		
Fully restricted PRB resources	O		INTEGER	Total number of available PRBs which are fully restricted by transmitted power. The statistical format is as defined by C3 in the C3-INNITIATED ACTION REQUEST message.		
Medium restricted PRB resources	O		INTEGER	Total number of available PRBs which are medium restricted by transmitted power.		

				The statistical format is as defined by C3 in the C3-INNITIATED ACTION REQUEST message.		

A.3.1.3 Traffic requirements

A base station or a serving radio node can further combine the traffic requirements from different UEs before reporting them to C3. The traffic requirements can be reported as a total or per network slice.

The role of C3 is to analyze these requirements and to assign the suitable radio nodes for uplink and/or downlink communication for each mentioned address.

C3 will select the appropriate cell or, if carrier aggregation is used, cells, for serving this demanding traffic. The non-demanding control information may be identified by another address and routed through the same or a different radio node.

Table A.4 Traffic requirements

IE/Group Name	**Presence**	**Range**	**IE type and reference**	**Semantics description**	**Criticality**	**Assigned Criticality**
R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
Cell General Identifier in NR	M		Integer (1..4000000)	NCGI	YES	Reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
R-TP Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Start SFN	M		INTEGER (0..1023, …)	SFN of the radio frame containing the first subframe when the requirements are valid.	–	–
Start Subframe Number	M		INTEGER (0..9, …)	Subframe number, within the radio frame containing the first subframe when the requirements are valid.	–	–
End SFN	M		INTEGER	Last considered	YES	ignore

			(0..1023, …)	radio frame		
Total traffic requirements	M		INTEGER (in Kb/s)	Total traffic requested by the served UEs. The statistical format is as defined by C3 in the C3-INNITIATED ACTION REQUEST message.		
Traffic per first slice	O		INTEGER (in kb/s)	Total traffic requested for the first slice. The statistical format is as defined by C3 in the C3-INNITIATED ACTION REQUEST message.		
Traffic per 2nd slice	O		INTEGER (in kb/s)	Total traffic requested for the 2nd slice. The statistical format is as defined by C3 in the C3-INNITIATED ACTION REQUEST message.		
Traffic per 3d slice	O		INTEGER (in kb/s)	Total traffic requested for the 3d slice. The statistical format is as defined by C3 in the C3-INNITIATED ACTION REQUEST message.	–	–

A.3.1.4 Available DL throughput per cell

This IE provides for each cell served by an R-TP the available DL throuput computed over the free resources based on a declared MCS and transmission mode.

Table A.5 Available DL throughput per cell

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
R-TP Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Start SFN	M		INTEGER (0..1023, …)	SFN of the radio frame containing the first subframe	–	–
Start Subframe Number	M		INTEGER (0..9, …)	Subframe number, within the radio frame indicated by the Start SFN IE.	–	–
End SFN	M		INTEGER (0..1023, …)	Last considered radio frame	YES	ignore
R-TP Cell list to be considered		1				
>Cell Item		1 .. <maxn oof Cells>			EACH	ignore
>>Cell General Identifier in NR	M		Integer (1…4000000)	NCGI		
>> Used MCS	M		Integer (1…32)	Index of MCS used in evaluation		
>>Used transmission mode	M		Integer (1…32)	Index of transmission mode		
>>Available traffic per cell	O		INTEGER (in kb/s)	Total available traffic estimated when using the specified MCS and transmission mode. The statistical format is as defined by C3 in the C3-INNITIATED ACTION	–	–

				REQUEST message.		

A.3.1.5 Coupling loss

Details on the Coupling Loss can be found in the deliverable D3.1 [4].

Table A.6 Coupling loss

This IE provides the coupling loss between an R-TP and an UE. The result is provided using the indicated statistical processing.IE/Group Name	**Presence**	**Range**	**IE type and reference**	**Semantics description**	**Criticality**	**Assigned Criticality**
R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
Cell General Identifier in NR	M		Integer (1..400000)	NCGI	YES	Reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
R-TP Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
R-TP Cell list to be considered		1				
>Cell Item		1 .. <maxnoof Cells>			EACH	ignore
>>Global_Cell_ID	M		Integer (1…4000000)	Value of Global_Cell_ID		
>>UE_F1_ID	M		Integer	UE_F1_ID		
>> Coupling loss	M		Integer (1…32)	Coupling loss in linear form; the statistical format is as defined by C3 in the C3-INNITIATED		

				ACTION REQUEST message.		

A.3.2 Information elements for control/coordination over F1 interface (described in D2.2)

The IEs in this section are sent by C3/CU to R-TP/DU.

A.3.2.1 Allocation of DL Time-Frequency Resources

This IE provides the maximum allocation for DL transmission; the allocation is valid until the next change.

Table A.7 Allocation of DL Time-Frequency Resources

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
C3 Control ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Start SFN	M		INTEGER (0..1023, …)	SFN of the radio frame containing the first subframe when the Time, frequency, and space resources per cell are valid.	–	–
Cell Group per R-TP		(1…< MaxCellsPerR-TP)				
>Cell Item						
>>Cell General Identifier in NR	M		Integer (1…4000000)	NCGI	YES	Reject
>>Start Subframe Number	M		INTEGER (0..9, …)	Subframe number, within the radio frame indicated by the Start SFN IE.	–	–
>>Subframe Assignment to R-TP in DL (Time resources for DL)	M		BITSTRING (SIZE(32))	Each position in the bitmap indicates a slot, “”1” indicates that the slot can be used for DL	YES	ignore

				transmission. First bit indicates the first slot, 2nd bit the second slot and so on. The report uses the report format in the initial C3 RESOURCE STATUS REQUEST message.		
>>Pattern repetition interval	O		INTEGER	Indicated as number of subframes	-	-
>>Number Of Cell-specific Antenna Ports	M		ENUMERATED (1, 2, 4, …)	P (number of antenna ports for cell-specific reference signals) t.b.d. in TS 38.211	–	–
>>P_B	M		INTEGER (0..3, …)	t.b.d in TS 38.213	–	–

A.3.2.2 Minimum traffic to be served per network slice

When assigning resources to each network slice, the base station or the C3 should make sure that all the served UEs have assigned at least the minimum requested resources for operation such to make sure that the minimum traffic is served. In this mode will be satisfied the operator requirement of isolating the network slices one from the other.

Table A.8 Minimum traffic to be served per network slice

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
Cell General Identifier in NR	M		Integer (1…4000000)	NCGI	YES	Reject
C3 Control ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Slice Group Items		(1…<MaxSlicesPerR-TP)				
>Slice item						
>>Start SFN	M		INTEGER (0..1023, …)	SFN of the radio frame containing the first subframe when the requirements are valid.	–	–
>>Statistical meaning	O		BITSTRING (SIZE(8))	Each position in the bitmap indicates the processing which the R-TP is requested to use for the control interval. First bit indicates average value, 2nd bit indicates minimum value. Remaining bits:"0".	–	–
Minimum traffic per slice	O		INTEGER (in kb/s)	Minimum traffic to be provided for the slice.	–	–

A.3.2.3 Functional Partition between the R-TP and the vRP

Depending of R-TP capabilities, the following situations can be implemented:

- A. PDPC - Virtualised;
- B. PDPC and RLC virtualised;
- C. PDPC, RLC, MAC virtualised;
- D. PDPC, RLC, MAC, and virtualisation of one of more of: CRC adding to TB, Code block segmentation and code block CRC attachment, Channel coding, Rate matching, Code block concatenation, scrambling.

- E. None (i.e. no VM used).

Table A.9 Functional Partition between the R-TP and the vR

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
C3 Control ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Start SFN	M		INTEGER (0..1023, …)	SFN of the radio frame containing the first sub-frame when the requirements are valid.	YES	reject
Functions to be executed in gNB-CU	M		BITSTRING (SIZE(8))	Each position in the bitmap indicates the processing which the gNB-CU will accomplish. Bit0=”1” indicates “PDCP”, bit1=”1” indicates “RLC”, bit 2=”1” indicates “MAC”, bit3=”1” indicates “all PHY functions”, bit4=”1” indicates “partial PHY functions”.	–	–
Partial PHY functions to be executed in gNB-CU	O		BITSTRING (SIZE(8))	Each position in the bitmap indicates the PHY processing which the gNB-CU will accomplish. Bit0=”1” indicates “CRC adding to TB”, bit1=”1” indicates “Code block segmentation and CRC attachment”, bit 2=”1” indicates “Channel coding”, bit3=”1” indicates “Rate matching,”, bit4=”1” indicates “Code block concatenation”, bit4=”1” indicates “scrambling”.		

A.3.2.4 Multi-point transmission

C3 allocates the time-frequency resources per R-TP. The UE will receive in each sub-set of slots data from a given R-TP. In other words, the R-TP provides support for a sub-cell.

The coding is identical with the coding for "Allocation of DL Time-Frequency Resources".

A.3.2.5 Handover

C3 will establish which are the new target R-TPs and will send a message to the source R-TP for informing the UE to which new cells shall connect.

Table A.10 Multi-point transmission

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Source R-TP_ID	M		Integer (1…4096...)	R-TP (Distributed Unit) ID	YES	Reject
C3 Control ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Group R-TP		(1…<maxServing R-TPs)				
>RTP Item						
>>Start SFN	M		INTEGER (0..1023, …)	SFN of the radio for starting the HO process	YES	reject
>>Cell Group per R-TP		(1…<MaxCellsPer R-TP)				
>>>Cell Item						
>>>>Cell General Identifier in NR	M		Integer (1…4000000)	NCGI	YES	Reject

A.3.2.6 Load balancing

The motivation for Load Balancing is different from the motivation for HO, however it uses the HO procedure, such that it can be used the IE defined for HO.

A.4 Information elements for the heterogeneous WiFi-LTE network

A.4.1 Reports

A.4.1.1 Received power histogram in energy detection

Listening node will just sample the received power and memorize it together with a time stamp, creating a histogram.

In case of stand-alone AP, eNB, gNB, R-TP ID represents the ID of AP, eNB, gNB.

A similar message can be created for STA.

The measurement duration is limited to 10min.

Table A.11 Received power histogram in energy detection

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Node_ID	M		Integer (1…4096...)	Node_ID (Distributed Unit or stand-alone AP, gNB, eNB) ID	YES	Reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Node Measurement ID	M		INTEGER (1..4095,...)	Allocated by the Node	YES	reject
Choice Start time						
>Start time in SFN						
>>			INTEGER (1…60000)	SFN	–	–
>Start time as absolute time						
>Start time			t.b.c	Absolute time (hour:min:sec:ms)	–	–
Group power-time samples for ED	M	1			–	–
>Item power-time samples for ED	O	(1…<maxpower-time-elements>)			–	–
>>Power sample in ED	O		INTEGER (0….1023)	Power in dB above -100dBm		
>>Relative Time	O		INTEGER	Time in microseconds relative to start time		

A.4.1.2 Power usage indication

Histogram of transmission power per node and channel

Table A.12 Power usage indication

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Node ID	M		Integer (1…4096...)	Node_ID (Distributed Unit or stand-alone AP, gNB, eNB) ID	YES	Reject
Channel ID	M		Integer (1…4096...)	Lowest channel ID	YES	Reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Node BW	M		INTEGER (1..10000,...)	BW in MHz	YES	reject
Node Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Choice Start time						
>Start time in SFN						
>>			INTEGER (1…60000)	SFN	–	–
>Start time as absolute time						
>Start time			t.b.c	Absolute time (hour:min:sec:ms)	–	–
Group power-time samples per transmission	M	1			–	–
>Item Tx power-time samples	O	(1…<maxpower-time-elements>)				
>> Power	O		INTEGER (0….1023)	Power in dB above -20dBm.		
>>Relative Time	O		INTEGER	Time in microseconds relative to start time		

A.4.1.3 RSSI log file and statistical information in active state

Log file containing statistical information of the power measured as RSSI of past receptions of a s

Table A.13 RSSI log file and statistical information in active state

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Node ID	M		Integer (1…4096...)	Node_ID (Distributed Unit or stand-alone AP, gNB, eNB) ID	YES	Reject
Channel ID	M		Integer (1…4096...)	Lowest channel ID	YES	Reject
Node BW	M		INTEGER (1..10000,...)	BW in MHz	YES	reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Node Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Choice Start time						
>Start time in SFN						
>>			INTEGER (1…60000)	SFN	–	–
>Start time as absolute time						
>Start time			t.b.c	Absolute time (hour:min:sec:ms)	–	–
Group RSSI samples per signal reception	M	1			–	–
>Item Rx power-time samples	O	(1…<maxpower-time-elements>)				
>> Rx Desired Signal Power	O		INTEGER (0….1023)	Power in dB above -100dBm. The report uses the report format in the initial C3 RESOURCE STATUS REQUEST message.		

>>Relative Time	O		INTEGER	Time in microseconds relative to start time		
>>Measurement Duration Time	O		INTEGER	Time in microseconds	–	–

A.4.1.4 RSSI log file and statistical information in idle state

Log file containing statistical information of the power measured as RSSI of past idle state of specific node on the operating channel.

Table A.14 RSSI log file and statistical information in idle state

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Node ID	M		Integer (1…4096...)	Node_ID (Distributed Unit or stand-alone AP, gNB, eNB) ID	YES	Reject
Channel ID	M		Integer (1…4096...)	Lowest channel ID	YES	Reject
Node BW	M		INTEGER (1..10000,...)	BW in MHz	YES	reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Node Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Choice Start time						
>Start time in SFN						
>>			INTEGER (1…60000)	SFN	–	–
>Start time as absolute time						
>Start time			t.b.c	Absolute time (hour:min:sec:ms)	–	–
Group RSSI samples	M	1			–	–
>Item Rx power-time samples	O	(1…<max power-time-elements>)				
>> Rx Idle Power	O		INTEGER (0….1023)	Power in dB above -100dBm. The report uses the report format in the initial C3 RESOURCE STATUS REQUEST message.		
>>Relative Time	O		INTEGER	Time in microseconds relative to start time		
>>Measurement Duration Time	O		INTEGER	Time in microseconds	–	–

A.4.1.5 RSSI log file and statistical information in not-used channels

Log file containing statistical information of the power measured as RSSI of past receptions of a specific node in non-operating channel.

Table A.15 RSSI log file and statistical information in not-used channels

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Node ID	M		Integer (1…4096...)	Node_ID (Distributed Unit or stand-alone AP, gNB, eNB) ID	YES	Reject
Channel ID	M		Integer (1…4096...)	Lowest channel ID	YES	Reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Measurement BW	M		INTEGER (1..10000,...)	BW in MHz	YES	reject
Node Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Choice Start time						
>Start time in SFN						
>>			INTEGER (1…60000)	SFN	–	–
>Start time as absolute time						
>Start time			t.b.c	Absolute time (hour:min:sec:ms)	–	–
Group RSSI samples per signal reception	M	1			–	–
>Item Rx power-time samples	O	(1…<maxpower-time-elements>)				
>> Rx Signal Power	O		INTEGER (0….1023)	Power in dB above -100dBm. The report uses the report format in the initial C3 RESOURCE STATUS REQUEST message.		
>>Relative Measurement Start Time	O		INTEGER	Time in microseconds relative to start time		
>>Measurement Duration Time	O		INTEGER	Time in microseconds	–	–

A.4.2 C3 actions for QoS enfoncement

A.4.2.1 Medium access parameters

Control of operational channel of a node and medium access parameters: LBT carrier sense power, LBT energy threshold, maximum medium occupancy duration.

Each channel can use different parameters.

Table A.16 Medium access parameters

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Node ID	M		Integer (1…4096...)	Node_ID (Distributed Unit or stand-alone AP, gNB, eNB) ID	YES	Reject
C3 Control ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Choice Allocation Start time						
>Start time in SFN						
>>Start SFN			INTEGER (1…60000)	SFN	–	–
>Start time as absolute time						
>>Start time			t.b.c	Absolute time (hour:min:sec:ms)	–	–
Group Assigned operational channels	M	1			–	–
>Item Assigned operational channel		(1…<channelsper node)				
>> Channel ID	M		Integer (1…4096...)	Channel ID	-	-
>>LBT lowest threshold	O		INTEGER	Power in dB relative to -100dBm		
>>LBT ED threshold	O		INTEGER	Power in dB relative to -100dBm	–	–
>>Maximum Occupancy Duration	O		INTEGER	Time in microseconds		

A.5 RESOURCE STATUS REPORTING

This section provides the semantics for the messages described in section

A.5.1 RESOURCE STATUS REQUEST

This message is sent by C3 (CU) to served R-TPs (DUs or gNBs or eNBs) to initiate the requested measurement according to the parameters given in the message.

Report Content IE indicates which reports shall be provided.

Report Format IE indicates the measurement processing which the gNB-DU is requested to use for the reporting interval. The pre-processing includes: average value, maximum value, minimum value, median value, 10th percentile, 90th percentile in the CDF.

Cell ID IE indicates for which cell shall be provided the reports.

Reporting Periodicity IE indicates the interval between reports.

Direction: C3 → R-TP.

Table A.17 RESOURCE STATUS REQUEST

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Message Type	M		6.2.a		YES	reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3 (or CU)	YES	reject
Report Content	O		BITSTRING (SIZE(32))	Each position in the bitmap indicates the measurement which the R-TP is requested to report. The measurements are indicated in D3.2 section 6.2.3. First bit indicates the first listed measurement, 2nd bit the second listed measurement and so on.	YES	reject
Report format	O		BITSTRING (SIZE(16))	Each position in the bitmap indicates the measurement processing which the R-TP is requested to use for the reporting interval. First bit indicates average value, 2nd bit indicates maximum value, 3d bit indicates minimum value, 4th bit indicates median value, 5th bit indicates 10th percentile, 6th bit indicates 90th percentile in the CDF.	YES	reject
Cell To Report		1		Cell ID list to which the request applies.	YES	ignore

>Cell To Report Item		1 .. <maxCellsinR-TP >			EACH	ignore
>>Cell ID	M		NCGI (global cell identifiers in NR)		–	–
Reporting Periodicity	O		ENUMERATED(50 ms, 100ms, 200ms, 500ms, 1000ms, 2000ms, 5000ms,10000ms, …)	Periodicity that can be used for reporting	YES	ignore

Range bound	**Explanation**
maxCellingNB	Maximum no. cells that can be served by a gNB. Value is 4000000.

A.5.2 RESOURCE STATUS RESPONSE

This message is sent by an R-TP to indicate that the requested measurement, for all or for a subset of the measurement objects included in the measurement request is successfully initiated.

In case that a measurement fails to be initialized, the message identifies such measurement by *Failed Report Characteristics* IE and indicates the reason for this failure by using the *Cause* IE.

Direction: R-TP → C3.

Table A.18 RESOURCE STATUS RESPONSE

IE/Group Name	**Pres-ence**	**Range**	**IE type and reference**	**Semantics description**	**Critic-ality**	**Assigned Criticality**
Message Type	M		6.2.(a+1)		YES	reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
R-TP Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Criticality Diagnostics	O		9.2.7 in TS36.423		YES	ignore

Measurement Initiation Result		0..1		List of all cells in which measurement objects were requested, included when indicating partial success	YES	ignore
>Measurement Initiation Result Item		1 .. <maxCellineNB>			EACH	ignore
>>Cell ID	M		NCGI		–	–
>>Measurement Failure Cause List		0..1		Indicates that R-TP could not initiate the measurement for at least one of the requested measurement objects in the cell	–	–
>>>Measurement Failure Cause Item		1 .. <maxFailedMeasObjects>			EACH	ignore
>>>>Measurement Failed Report Characteristics	M		BITSTRING (SIZE(32))	Each position in the bitmap indicates the measurement which the R-TP was requested to report. The measurements are indicated in D3.2 section 6.2.3. First bit indicates the first listed measurement, 2nd bit the second listed measurement and so on.	–	–
>>>>Cause	M		t.b.c	Failure cause for measurement objects for which the measurement cannot be initiated	–	–

Range bound	**Explanation**
maxFailedMeasObjects	Maximum number of measurement objects that can fail per measurement. Value is [32].
maxCellInGNB	Maximum no. cells that can be served by a gNB. Value is 4000000.

A.5.3 RESOURCE STATUS FAILURE

This message is sent by the R-TP to indicate that for none of the requested measurement objects the measurement can be initiated.

Direction: R-TP → C3.

Table A.19 RESOURCE STATUS FAILURE

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Message Type	M		6.2.(a+2)		YES	reject
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
R-TP Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Cause	M		t.b.c.	Ignored by the receiver when the Complete Failure Cause Information IE is included	YES	ignore
Criticality Diagnostics	O		t.b.c		YES	ignore
Complete Failure Cause Information		0..1		Complete list of failure causes for all requested cells.	YES	ignore
a. >Complete Failure Cause Information Item		1 .. <maxCellineNB>			EACH	ignore
>>Cell ID	M		NCGI		–	–
>>Measurement Failure Cause List		1			–	–
>>>Measurement Failure Cause Item		1 .. <maxFailedMeasObjects>			EACH	ignore

>>>>Measurement Failed Report Characteristics	M		BITSTRING (SIZE(32))	Each position in the bitmap indicates measurement object that failed to be initiated in the R-TP. Each position in the bitmap indicates the measurement which the R-TP was requested to report. The measurements are indicated in D3.2 section 6.2.3. First bit indicates the first listed measurement, 2nd bit the second listed measurement and so on.	–	–
>>>>Cause	M		9.2.6	Failure cause for measurements that cannot be initiated	–	–

Range bound	Explanation
maxCellineNB	Maximum no. cells that can be served by an eNB. Value is 256.
maxFailedMeasObjects	Max number of measurement objects that can fail per measurement. Value is 32.

A.5.4 RESOURCE STATUS UPDATE

This message is sent by R-TP to C3 to report the results of the requested measurements.

Direction: R-TP → C3.

Table A.20 RESOURCE STATUS UPDATE

Range bound	Explanation
maxCellineNB	Maximum no. cells that can be served by an eNB. Value is 256.

Condition	Explanation
ifRegistrationRequestStoporPartialStoporAdd	This IE shall be present if the Registration Request IE is set to the value “stop”, “partial stop” or “add”.

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Message Type	M		6.2.(a+3)		YES	ignore
C3 Measurement ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
R-TP Measurement ID	M		INTEGER (1..4095,...)	Allocated by R-TP	YES	reject
Cell Measurement Result		1			YES	ignore
>Cell Measurement Result Item		1 .. <maxCellineNB>			EACH	ignore
>>Cell ID	M		NCGI			
>>Time, frequency and space resources per cell in downlink	O		Annex			
>>summary of resource availability per cell	O		Annex		YES	ignore
>>Traffic requirements	O		Annex		YES	ignore
>>Coupling loss in DL reported by UE	O		Annex		YES	ignore

A.6 C3-INNITIATED ACTION

A.6.1 C3-INNITIATED ACTION REQUEST

The generic form of this message is presented below.

Direction: C3 -> R-TP.

Table A.21 C3-INNITIATED ACTION REQUEST

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
Message Type	M		6.2.(a+4)		YES	ignore
C3 Control ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
C3 Control Group		1			YES	ignore
>C3 Control Item		1 .. <maxCellineNB>			EACH	ignore
>>Cell ID	M		NCGI			
>> Allocation of DL Time-Frequency Resources	O		Annex			
>> Minimum traffic to be served per network slice	O		Annex		YES	ignore
>>Functional Partition between the R-TP and the vRP	O		Annex		YES	ignore
>>R-TP handover target	O		Annex		YES	ignore
>> Load balancing targets	O		Annex		YES	ignore

A.6.2 C3-INNITIATED ACTION RESPONSE

A message will be sent by R-TP to indicate if the C3 request can be executed or not.

A.6.3 C3-INNITIATED ACTION FAILURE

R-TP provides to C3 the cause of failure.

A.7 CONTROL OF HETEROGENEOUS IEEE 802.11– 3GPP LAA NETWORK

A.7.1 General Reports

A.7.1.1 Protected resources in LE (License-exempt) bands

A protected resource in a LE band is a time-limited resource that, after determining that the medium is free, is not accessible for more than one transmitter at any given time. There may be different power criteria for assessing if the medium is free or occupied.

Listen before talk (LBT) in WiFi, used also in LTE mode named licensed assisted access (LAA), is also a form of protection, given that only one transmitter on a channel will use the channel when the received signal is higher than a threshold. In case that such a transmission takes place, the other potential transmitters shall use the resource in a fully restricted mode. The available resources is the time percentage, calculated based on past medium sensing, in which the level of energy is under the LBT threshold.

Specific message parameters:

- General cell ID;
- Time interval for protection
- LBT threshold to be applied.

A.7.1.2 Energy consumption

Traffic scheduling should be done in such a way to minimize the overall energy consumption. C3 needs to consider the overall consumed energy; for example, blanking resources will reduce the energy consumption of both the transmitting node, as can use higher MCS, and of the blanking node, which will transmit zero power. This energy reduction will be done by spending more spectrum, which in practice may not be always possible.

For an UE in the active state the transmission-related energy consumption, excluding the energy required by the general computing, will depend mainly on the radio used power and the time of transmission.

For an UE the relevant parameters are:

- Total transmission power on all the used antennas;
- Time of transmission.

So for uplink should be selected serving radio nodes with the lowest path loss or coupling loss and time-frequency resources less affected by interference. The serving eNB can schedule the actual resources for transmission, while C3 can coordinate resources to be used in Network MIMO or multi-connectivity; in the last case is not needed that the transmissions to different radio nodes will be executed in the same time. For the network to UE transmission there are more parameters that can influence the receiving radio node selection and the appropriate RAT.

For example in a heterogeneous deployment the macro base station signal may be stronger than a small cell signal, however the energy cost for transmitting the same throughput can be much higher for a macro base station due to:

- Air conditioning
- Losses in cables between the base band enclosure and the antennas
- Higher path loss to UE.

For providing an energy-efficient solution, the radio-node selection by C3 should consider parameters provided through a message by the base station or by the radio node including:

- Background (due to computing and air conditioning) consumed power
- Consumed power for each additional W of radio transmission.
- Consumed power for powering-up the radio node.

The base station or C3 should assess, for each potential radio transmission node, the needed transmit power based on:

- Transmission power of the reference signals
- UE measurement reports (RSRP, RSRQ, CQI, MIMO rank, etc.) or based on the network graph assessing the CQI and MIMO rank.
- Cost of powering up a radio node.

In addition, different RATs may have different peak-to-average properties which influence the consumed power, so the assessment should be done per RAT. Also the operating frequency influences the transmission power, such that the evaluation should take into consideration the suitable cells. Of course, the smooth mobility and coverage requirements may justify spending more energy, such that this aspect should be also considered when selecting the radio transmitting or receiving nodes.

A.8 C3 interaction with user plane

A.8.1 Downlink bearer re-direction for HO

This message is used in the disaggregated architecture, where the C3 and the user plane router are not co-located. For fast operation, C3 is setting the target TEID from a pool of TEIDs in its control.

Table A.22 Downlink bearer re-direction for HO

IE/Group Name	Presence	Range	IE type and reference	Semantics description	Criticality	Assigned Criticality
C3 Control ID	M		INTEGER (1..4095,...)	Allocated by C3	YES	reject
Group R-TP		(1…<maxServingR-TPs)				
>RTP Item						
>>Source Tunnel End Point ID	O		INTEGER	TEID at R-TP	YES	Reject
>> Target R-TP	O		INTEGER	Target R-TP ID	YES	Reject
>> Target Temporary R-TP Tunnel End Point ID				From the TEID pool allocated by C3		
>>Start SFN for Bearer re-direction	O		INTEGER	SFN of the radio for starting the HO process	YES	reject