iJOIN

INFSO-ICT-317941

INFSO-ICT-317941 iJOIN

D3.2

Definition of MAC and RRM approaches for RANaaS and a joint backhaul/access design

Editor:	Emmanouil Pateromichelakis, UNIS
Deliverable nature:	Deliverable
Suggested readers:	iJOIN GA
Due date:	October 31th, 2014
Delivery date:	October 31th, 2014
Version:	1.0
Total number of pages:	124
Reviewed by:	GA_members
Keywords:	MAC/RRM State-of-the-art, RAN MAC, Backhaul MAC, RANaaS, Functional Split, iJOIN
Resources consumed	61.92 PM

Abstract

This deliverable defines the set of MAC and RRM candidate technologies considered by the iJOIN architecture. It particularly shows how they integrate in the iJOIN architecture, how they contribute to iJOIN's objectives and which practical constraints and requirements are considered. This deliverable serves as input for the final proof-of-concept and provides a harmonised view with other work packages.

List of authors

Company	Author
CEA	Antonio De Domenico (antonio.de-domenico@cea.fr)
	Dimitri Ktenas (dimitri.ktenas@cea.fr)
HP	Marco Di Girolamo (marco.digirolamo@hp.com)
	Beppe Coffano (beppe.coffano@hp.com)
	Marco Consonni (marco.consonni@hp.com)
IMC	Umer Salim (umer.salim@intel.com)
IMDEA	Jorge Ortín (jortin@it.uc3m.es)
NEC	Andreas Maeder (andreas.maeder@neclab.eu)
	Peter Rost (peter.rost@neclab.eu)
SCBB	Massinissa Lalam (massinissa.lalam@sagemcom.com)
UoB	Henning Paul (paul@ant.uni-bremen.de)
UNIS	Emmanouil Pateromichelakis (e.pateromichelakis@surrey.ac.uk)
TUD	Richard Fritzsche (richard.fritzsche@tu-dresden.de)

History

Modified by	Date	Version	Comments
Emmanouil Pateromichelakis	October 31st 2014	1.0	Final version of D3.2

Table of Contents

List of authors ... 2
History ... 3
Table of Contents ... 4
List of Figures ... 6
List of Tables ... 8
Abbreviations ... 9
1 Executive Summary ... 12
2 Introduction ... 13
2.1 Motivation and Background ... 13
2.2 Key Contributions ... 13
3 Functional Split and veNB ... 16
3.1 iJOIN Architecture ... 16
3.2 Functional Split ... 17
3.2.1 Functional Split Decision Factors ... 18
3.2.1.1 Backhaul constraints ... 18
3.2.1.2 3GPP requirements ... 21
3.2.1.3 Minimal centralization requirements of CTs ... 25
3.2.1.4 Lower layer dependencies ... 26
3.2.2 Functional Split Options ... 27
3.2.3 Functional Split Configurations ... 30
3.3 Virtual eNB Implementation ... 31
3.3.1 CT Virtualization ... 31
3.3.2 Functional constraints ... 32
4 iJOIN MAC/RRM Candidate Technologies ... 38
4.1 CT 3.1: Backhaul Link Scheduling and QoS-aware Flow Forwarding ... 38
4.1.1 Technical description ... 38
4.1.2 Implementation of CT in the iJOIN architecture ... 41
4.1.3 Evaluation of the CT ... 42
4.2 CT 3.2: Partly decentralized mechanisms for joint RAN and backhaul optimization in dense small cell deployments ... 44
4.2.1 Technical description ... 44
4.2.2 Implementation of CT in the iJOIN architecture ... 50
4.2.3 Evaluation of the CT ... 51
4.3 CT 3.3: Energy-Efficient MAC/RRM at Access and Backhaul ... 54
4.3.1 Technical description ... 54
4.3.2 Implementation of CT in the iJOIN architecture ... 57
4.3.3 Evaluation of the CT ... 58
4.4 CT 3.4: Computational Complexity and Semi-Deterministic Scheduling ... 59

4.4.1 Technical description ... 59
4.4.2 Implementation of CT in the iJOIN architecture ... 61
4.4.3 Evaluation of the CT ... 61
4.5 CT 3.5: Cooperative RRM for Inter-Cell Interference Coordination in RANaaS ... 64
4.5.1 Technical description ... 64
4.5.2 Implementation of CT in the iJOIN architecture ... 67
4.5.3 Evaluation of the CT ... 67
4.6 CT 3.6: Utilization and Energy Efficiency ... 69
4.6.1 Technical description ... 69
4.6.2 Evaluation of the CT ... 72
4.7 CT 3.7: Radio Resource Management for Scalable Multi-Point Turbo Detection ... 75
4.7.1 Technical description ... 75
4.7.2 Implementation of CT in the iJOIN architecture ... 80
4.7.3 Evaluation of the CT ... 84
4.8 CT 3.8: Radio Resource Management for In-Network-Processing ... 86
4.8.1 Technical description ... 86
4.8.2 Implementation of CT in the iJOIN architecture ... 87
4.8.3 Evaluation of the CT ... 87
4.9 CT 3.9: Hybrid local-cloud-based user scheduling for interference control ... 88
4.9.1 Technical description ... 88
4.9.2 Implementation in the iJOIN architecture ... 91
4.9.3 Evaluation of the CT ... 91
5 Overall Evaluation ... 94
6 Summary and Conclusion ... 96
Acknowledgements and Disclaimer ... 97
Appendix I Input and Output Parameters ... 98
Appendix II Evaluation methodology ... 103
II.1 Radio Access Network Modelling ... 103
II.1.1 Outdoor small cell deployment ... 103
II.1.2 Indoor small cell deployment ... 108
II.2 iJOIN evaluation scenarios ... 110
II.2.1 Outdoor deployment ... 111
II.2.2 Indoor deployment ... 112
Appendix III Categorization of Backhaul Technologies ... 113
Appendix IV CT interactions in WP3 ... 114
References ... 121

List of Figures

Figure 3-1: WP3 functional architecture. 17

Figure 3-2: Break down of timing constraints for centralized HARQ in LTE FDD with < 1 ms one-way backhaul latency and ~ 2ms RANaaS processing and frame building delay. 22

Figure 3-3: Impact of latency on hand-over failure rate (HOF) 24

Figure 3-4: Mapping of the functional split scenario to the protocol stack for CRA 26

Figure 3-5: Illustration of Centralized Connection Control 26

Figure 3-6: Functional split options on MAC layer 27

Figure 3-7: Dual connectivity as developed by 3GPP [61] 28

Figure 3-8: X2-C interface [61] 28

Figure 3-9: Example functional split configuration for CRA 30

Figure 3-10: Illustration of CCC in case of centralized PHY and MAC processing 31

Figure 3-11: Server Virtualization 33

Figure 3-12: Interrupt Management in a virtualized environment 34

Figure 3-13: Non-optimized vs. optimizated virtualized MSI latency according to [67] 35

Figure 3-14: Mapping of CTs to virtual machines 36

Figure 4-1: Backhaul link scheduling and QoS-aware flow forwarding 38

Figure 4-2 Back-pressure scheduling example for 4 iSCs 41

Figure 4-3 Message Sequence Chart for CT3.1 42

Figure 4-4 Small Cell Deployment for CT3.1 42

Figure 4-5 Illustration of BH topology for k=19 and k=1 43

Figure 4-6 Average BH link spectral efficiency vs. number of routes 43

Figure 4-7 Maximum and average delay for different delay bounds vs. number of routes 44

Figure 4-8 BH link Spectral Efficiency vs. Maximum Delay 44

Figure 4-9: The heterogeneous network deployment under investigation.F1 and F2 are the carrier frequencies for the macro layer and the small cell layer, respectively. 45

Figure 4-10: Evolve paradigm for managing user-eNB association. 48

Figure 4-11: Required message passing and functions in the centralized algorithm proposed in CT3.2. 50

Figure 4-12: Required message passing and functions in the centralized algorithm proposed in CT3.2. 51

Figure 4-13: Cumulative distribution function of the Network Shannon Capacity achieved with different association schemes. 52

Figure 4-14 CDF of the proposed solution vs benchmark 54

Figure 4-15 Comparison of the network capacity achieved with the different games 54

Figure 4-16: Comparison of the network log-sum rate achieved with the different games 54

Figure 4-17: Message sequence chart for CT 3.3 57

Figure 4-18: Cumulative Network EE with respect to different small cell management scheme. 58

Figure 4-19: Semi-deterministic, hierarchical scheduling. 59

Figure 4-20: Message sequence chart for semi-deterministic scheduling in CT3.4. 61

Figure 4-21: PDF of the known channel for different amplitudes 62

Figure 4-22: Outage probability as a function of the allocated rate 62
Figure 4-23: Outage probability as a function of the allocated rate amplitudes for an SNR of 5 dB. 63
Figure 4-24: Outage probability as a function of the allocated rate for an SNR of 15 dB. 63
Figure 4-25: Net rate as a function of the allocated rate for an SNR of 5 dB. 63
Figure 4-26: Net rate as a function of the allocated rate for an SNR of 15 dB. 63
Figure 4-27: Inter-cell interference coordination between iSC 64
Figure 4-28 Message sequence chart for Cooperative RRM as proposed in CT3.5 67
Figure 4-29 CDF of DL SINR (left) and CDF of cell spectral efficiency (right) 68
Figure 4-30 Cell Throughput Comparison for random vs. regular deployment 69
Figure 4-31: Utilization gains in different network domains 70
Figure 4-32: Computational effort as a function of the SINR 72
Figure 4-33: Computational utilization efficiency 73
Figure 4-34: SINR distribution in a HetNet scenario 73
Figure 4-35: Distribution of per-subframe normalized computational complexities 74
Figure 4-36: Trace of total computational complexity and number of attached UEs per cell (macro cells) 74
Figure 4-37: Per-cell computational complexity vs. number of attached UEs 75
Figure 4-38: Scalable multi-point turbo detection scenario (solid lines are minimal requirements) 76
Figure 4-39: CDF of the average UE UL throughput (a) and transmit power (b) with (red) and without (black) inner loop power control 79
Figure 4-40: MPTD functional split when involving UE2/iSC1 and UE3/iSC2 81
Figure 4-41: Message sequence chart for centralised RRM in case of MPTD 82
Figure 4-42: SPTD functional split when processing involving UE2/iSC1 and UE3/iSC2 83
Figure 4-43: Message sequence chart for centralised RRM in case of SPTD 83
Figure 4-44: Comparison of the small-cell uplink throughput CDF 85
Figure 4-45: Comparison of the user uplink throughput CDF 85
Figure 4-46: Comparison of the “paired” user uplink throughput CDF 86
Figure 4-47: Message sequence chart for INP-aware central RRM as proposed by CT3.8 87
Figure 4-48: Exemplary physical backhaul topologies for 4 iSCs, a) Point-to-Point, b) Point-to-Multi-Point, c) Point-to-Point with central iTN 88
Figure 4-49: Average throughput for different INP variants and central processing 88
Figure 4-50 Schematic presentation of the architecture as studied in CT3.9 89
Figure 4-51: Cooperative power control with distributed CSI at the iSCs. 89
Figure 4-52: Message sequence chart for the hybrid scheduling algorithm as proposed by CT3.9 91
Figure 4-53: Ergodic rate achieved with the different scheduling strategies for the pathloss parameters $\sigma_{11}^2=1,\sigma_{12}^2=1,\sigma_{21}^2=1,\sigma_{22}^2=1$. 92
Figure 4-54: Ergodic rate achieved with the different scheduling strategies for uniform pathloss with $K=5$. 93
Figure 6-1: Deployment scenarios of outdoor small cells with macro coverage [21]. 103
Figure 6-2: Deployment scenarios of outdoor smalls cell without macro coverage. 103
Figure 6-3: Deployment scenarios of indoor small cell without macro coverage [21]. 108

List of Tables

Table 3-1: iJOIN RRM/MAC Candidate Technologies (CTs) 16
Table 3-2: Backhaul Classification (based on [21] and D4.2 [59]) 21
Table 3-3: 3GPP timing requirements 22
Table 3-4: Mapping of configurations to CTs 25
Table 3-5: Comparison of split options 29
Table 3-6 Interrupt latency (based on [21]) 35
Table 4-1: $D_\alpha(s)^{\neg u}$ and $D_\alpha(s)^{\oplus u}$ with respect to different resource allocation policies. *η(u,s)* represents the spectral efficiency between u and s. 47
Table 4-2: System Level Simulation Static Parameters 76
Table 4-3: System Level Simulation Dynamic Parameters 78
Table 5-1: Qualitative impact of the WP3 CTs with respect to the global iJOIN objectives. 94
Table 5-2: Compatibility of the WP3 CTs 95
Table 6-1: Required Input of WP3 CTs 99
Table 6-2: Required Output of WP3 CTs 101
Table 6-3: List of Abbreviations 102
Table 6-4: 3GPP outdoor deployment assumptions [21]. 105
Table 6-5: Mapping of the iJOIN WP3 assumptions to the 3GPP outdoor model. 107
Table 6-6: ITU indoor deployment assumptions [21]. 109
Table 6-7: Mapping of the iJOIN WP3 assumptions to the ITU indoor model. 110
Table 6-8: iJOIN WP3 outdoor common evaluation scenario 111
Table 6-9: iJOIN WP3 indoor common evaluation scenario 112
Table 6-10: Backhaul Classification 113

Abbreviations

ARQ	Automatic Repeat Request
ASIC	Application Specific Integrated Circuit
AT	Area Throughput
BER	Bit Error Rate
BF	Brute Force
BH	Backhaul
CAS	Computationally aware Selection
CCC	Centralized Connection Control
CDF	Cumulative Distributed Function
CEff	Cost Efficiency
CN	Core Network
CoMP	Coordinated Multi Point
CP	Central Processor
CPU	Central Processing Unit
CRA	Coordinated Resource Allocation
CQI	Channel Quality Indicator
CSI	Channel State Information
CT	Candidate Technology
DHCP	Dynamic Host Configuration Protocol
DL	Downlink
DMRS	Demodulation Reference Signal
DSP	Digital Signal Processor
EEff	Energy Efficiency
eNB	evolved NodeB
FAPI	Femto Application Platform Interface
FDD	Frequency Domain Duplexing
FEC	Forward Error Coding
FPGA	Field Programmable Gate Array
FSO	Free Space Optical
FFT	Fast Fourier Transform
GRRC	Global Radio Resource Control
HARQ	Hybrid Automatic Repeat Request
IaaS	Infrastructure as a Service
ICIC	Inter-cell Interference Coordination
iJOIN	Interworking and JOINt Design of an Open Access and Backhaul Network Architecture for Small Cells based on Cloud Networks
iNC	iJOIN Network Controller
InH	Indoor/Hotspot
INP	In-Network-Processing
iSC	iJOIN Small Cell
iVeC	iJOIN Virtual eNB Controller
ITU-R	International Telecommunication Union – Radio
KVM	Kernel-based Virtual Machine
KPI	Key Performance Index
LoS	Line of Sight

LRRC	Local Radio Resource Control
LTE	Long Term Evolution
MAC	Medium Access Control
MCRA	Multi-Cluster Resource Allocation
MCS	Modulation and Coding Scheme
MDP	Markov Decision Process
MeNB	Macro eNB
MIESM	Mutual Information based exponential SNR Mapping
MMSE	Minimum Mean Square Error
mmW	millimetre wave
MPS	Minimum Path Selection
MPTD	Multi-Point Turbo Detection
MRS	Max-Rate Selection
MSC	Message Sequence Chart
MSI	Message Signal Interrupt
MUD	Multi User Detection
MUX	Multiplexing
NE	Nash Equilibrium
NLoS	Non Line of Sight
NP	Non-Polynomial
OFDMA	Orthogonal Frequency-Division Multiple Access
OS	Operating System
OSI	Open Systems Interconnection
PCI	Physical Cell Identity
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
PHICH	Physical HARQ Indicator Channel
PHY	Physical layer
PmP	Point-to-Multipoint
PoP	Point of Presence
PRB	Physical Resource Block
PtP	Point-to-Point
PUCCH	Physical Uplink Control Channel
PUSCH	Physical Uplink Shared Channel
RAM	Random Access Memory
RAN	Radio Access Network
RANaaS	RAN as a Service
RAT	Radio Access Technology
RB	Resource Block
RLC	Radio Link Control
RRC	Radio Resource Control
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
SDN	Software Defined Networking
SDU	Service Data Unit
SeNB	Secondary eNodeB

SINR	Signal-to-Interference-and-Noise Ratio
SNR	Signal-to-Noise Ratio
SON	Self Organizing Networks
SPTD	Single Point Turbo Detection
SRS	Sounding Reference Signal
TDD	Time Domain Duplexing
TTI	Transmission Time Interval
UE	User Equipment
UEff	Utilization Efficiency
UL	Uplink
vCPU	virtual CPU
vDSL	very high-bit-rate Digital Subscriber Line
veNB	virtual eNB
VM	Virtual Machine
VMM	Virtual Machine Monitor
Wi-Fi	Wireless Fidelity
WSRM	Weighted Sum-Rate Maximization

1 Executive Summary

This report describes the main activities carried out by WP3 during the M19-M24 period of the iJOIN project. The main objective of this report is to present preliminary results for joint access/backhaul radio resource management and a set of novel radio resource management algorithms for scenarios considering RANaaS centralization. The report is organized as follows:

Section 2 presents a brief introduction underlining the scope and the objectives of this deliverable. Moreover, the overall contributions as well as the key contributions per CT are highlighted in Section 2.1.

Section 3 elaborates how the virtual eNodeB (veNB) concept affects the WP3 activities, detailing the impact on the specific CTs and discusses the general concept of functional split. In addition to this, practical issues like the implementation of the RANaaS on a cloud infrastructure and the constraints imposed by compatibility with existing 3GPP standards are analysed. Moreover, the functional split options which are relevant to WP3, as well as functional split configuration examples, are thoroughly discussed.

Section 4 recalls the candidate technologies as defined in D3.1 [5], and presents the progress, including updates and preliminary results. Additionally, the methodology for the evaluation of the CTs is detailed and the functional split options, which determine the applicability of each CT for different iJOIN scenarios, are further analysed.

Section 5 discusses the evaluation outcomes of the CTs and presents a comparative study between some highlighted CTs to capture the impact of these contributions using common evaluation scenarios (based on 3GPP reference scenarios for WP3, as discussed in Appendix II).

This report is then summarized and concluded in Section 6.

This report also encloses four appendix sections. Appendix I summarizes the Input / Output parameters, initially defined in D3.1 [5], which are used in Section 4 for the description of each CT's implementation. Additionally, Appendix II discusses the evaluation methodology and the iJOIN common scenarios. Furthermore, Appendix III highlights the outcome of the discussion for the categorization of backhaul technologies, which was initiated by WP3. Finally, Appendix IV shows the detailed interaction of each CT in WP3 with the other CTs.

2 Introduction

2.1 Motivation and Background

The iJOIN project aims to design an enhanced mobile network architecture and system based on the two fundamental concepts of the RAN-as-a-Service (RANaaS) and the joint operation of the access and backhaul networks. Together it pursues the centralization of RAN functionality onto a common general purpose IT platform, and, benefiting from cloud computing concepts, proposing a flexible adaption of the RAN to a heterogeneous backhaul with varying properties. Within iJOIN, Work Package 3 (WP3) investigates Medium Access Control (MAC) / Radio Resource Management (RRM) solutions for the backhaul and access networks. These solutions are based on a holistic backhaul and access view considering very dense small-cell networks, leveraging also the RANaaS concept for improving flexibility and exploiting the cloud resources.

In this report, the concept of the virtual eNodeB (veNB) introduced in the previous deliverable [5] is elaborated in more detail. In particular, the implications of the veNB concept on the different CTs are investigated, elaborating also how the idea of the functional split between local and centralized processing can be aligned with the virtual eNB concept. In this context, the key indicators for the selection of the functional split are further analysed and the functional split options with regard to WP3 are explicitly described. Moreover, key aspects of the veNB implementation are discussed to better capture the impact of iJOIN architecture on the CT operation.

Furthermore, the set of CTs introduced in deliverable D3.1 [5] are assessed in depth and consequently refined. Their applicability to the iJOIN architecture defined in deliverable D5.1 [15] is outlined and by means of initial evaluation, it is analyzed how the iJOIN key objectives are addressed by the proposed approaches and how these are affected by practical constraints and requirements. In this direction, the implementation of each CT in the iJOIN architecture is discussed by means of explicit representation of the signalling required for different functional split options.

In order to ensure that the results between different partners and different CTs are comparable, a set of common simulation parameters for performing the CTs' evaluation is specified considering two distinct scenarios, one outdoor and one indoor.

2.2 Key Contributions

The overall project-wide contributions which are presented in this deliverable are the joint RAN/BH discussion and the definition of functional split options with regard to WP3 perspective. The key layer 2 functions are highlighted and the functional split is further de-composed to capture the effect of centralization for different functions. Initially, some key decision factors are defined (i.e. backhaul, 3gpp constraints, etc.), to help us identify which should be the best functional split to be selected. Based on these factors, the main layer-2 functional split options are introduced and some configuration examples are discussed as show cases. In this context, some veNB implementation aspects are further analysed, incorporating the per-CT virtualization and some functional constraints imposed by the RANaaS platform.

Regarding the individual CTs which are explicitly described in this deliverable, we also highlight below the key contributions which include updates of the CTs and novel results.

CT 3.1 investigates the joint path selection and backhaul link scheduling problem, taking into account millimetre wave backhaul between iSCs. Here, the objectives are firstly to dynamically identify links to be scheduled per time slot (by using the same routing tables as provided by upper layers) taking into account the target global objective for the network (in terms of maximizing backhaul capacity). Secondly, to identify how the incoming flows are stored in the queues and forwarded to the next hops, taking into account the link selections in the previous step and the fulfilment of the QoS requirements (delay, outage, data rate) per flow. This problem was decoupled in two sub-problems and was solved using Branch-and-Cut and Back-pressure scheduling. The results show the gain in performance by increasing adaptively the number of hops; however, this comes at the cost of higher delays which can be critical for certain types of traffic.

CT 3.2 proposes a novel cell selection method, where radio access and backhaul load are jointly considered. In this document, starting from the analysis presented in D3.1, we elaborate two heuristic mechanisms. The first one, named as Evolve, iteratively updates the set of UEs associated at each eNB, such that the overall

network capacity is optimized. Here, we discuss each step of the proposed algorithm and we present the associated functionalities. We also present an alternative heuristic solution based on game theory where each player aims to maximize a predefined utility function, such as the overall network capacity or the fairness.

In CT 3.3, we investigate optimal control of the small cell activity for energy saving purposes. The goal is to find an optimal policy that trade-offs energy saving and end user's QoS. In this deliverable, we model this problem as a Markov decision process in full observable environment. Preliminary results are presented highlighting the potential gains associated to the proposed mechanism.

CT 3.4 presents a robust proportional fair scheduling algorithm which takes into account that the CSI at the scheduler is impaired due to, e.g., feedback quantization or latency. The robust algorithm is able to either guarantee a fixed outage probability or to maximize the spectral efficiency. The scheme is applied to the iJOIN architecture, where the functional split is performed at the scheduler. Hence, the global resource allocation can be updated locally, based on less outdated channel knowledge, resulting in a multi-stage scheduling scheme.

CT 3.5 proposes a novel graph-based multi-cell scheduling framework at RANaaS for efficiently mitigating downlink inter-cell interference at iSCs. We start by transforming the conventional weighted sum-rate maximization problem into an equivalent graph-based optimization. Thereafter, we decompose the optimization problem into dynamic graph-partitioning based sub-problems across different sub-channels and provide an optimal solution using Branch-and-Cut approach. Subsequently, due to high complexity of the solution, we propose heuristic algorithms that display near optimal performance. At the final stage, we apply a cluster-based resource allocation per sub-channel to find candidate users with maximum total weighted sum-rate. A case study on networked iSCs is also presented with simulation results showing a significant improvement over the state-of-the-art multi-cell scheduling benchmarks in terms of outage probability as well as average cell throughput.

CT 3.6 investigates novel metrics for utilization and energy efficiency. The main challenge for both metrics is to develop a holistic view on the overall system including cloud resources for RANaaS, backhaul resources as well as RAN resources. For utilization efficiency, an approach based on weighted indices will be investigated, where the resources in each domain are normalized and included in an overall metric. The required cloud resources, which depend on the functional split configuration, are determined by means of an LTE turbo decoder implementation which constitutes the main contribution on computational demand. For energy efficiency, a similar principle is used where the computational demands determine the energy consumption in the RANaaS. Evaluation for utilization efficiency is performed with a 3GPP compliant system-level simulator.

CT 3.7 investigates the radio resource management (RRM) enabling the use of the multi-point turbo detection principle. This technique is meant to be applied in an uplink scenario, where edge users will be scheduled on the same resources. Iterative processing (turbo detection) exploits what was previously considered as interference to enhance the detection of all users. Ideally, such physical processing should be centralised (within RANaaS) but it could also be locally applied (in each iSC). For these two options, a centralised RRM is needed to associate the "right" small cells and the "right" users in order to benefit from this physical processing. This RRM algorithm runs on an on-demand basis, meaning that a central part is working in the RANaaS and that a local part is running in each involved iSC. The iSCs may also communicate with each other directly if necessary. We developed one single RRM algorithm, for both central and distributed processing, which identifies the users that could be paired together in the uplink based solely on long-term downlink measurements. To assess the performance on a large scale manner, we used uplink system-level simulations. The early results show that significant gain in terms of area throughput can be achieved, especially for centralised processing where the paired users experience better fairness (5-percentile being close to the average) and greater throughput.

CT 3.8 addresses the simulative analysis of the effect of different RRM mechanisms for use with the In-Network Processing (INP) technique for distributed multi-user detection (MUD) under practical backhaul constraints. INP enables the allocation of different users to the same physical resources through MUD facilitating iSC-to-iSC communication over J2 backhaul links, however, at the expense of J2 traffic and increased processing latency. Scope of this CT is the assessment of the trade-off between gain in area throughput and increased backhaul load. To this end, investigations on the achievable throughput have been jointly performed with its WP2 counterpart, [53].

CT 3.9 studies the problem of partially centralized distributed scheduling. The proposed approach consists in using the backhaul architecture to exchange the long-term statistical information of the multi-user channel. This long-term information is then used to develop a Bayesian scheduling scheme and hence enforce coordination between the scheduling decisions taken in a distributed manner on the basis of only local instantaneous CSI knowledge. The proposed approach is shown to achieve most of the performance improvement of coordinated scheduling at the cost of only small requirements in terms of backhaul and computation resources. More details on the proposed distributed scheduling algorithm can be found in [38].

The dissemination achievements in WP3 can be summarized below. In total 3 journal and 4 conference papers were published and many more papers have been submitted to prestigious journals and conferences.

In particular, in IEEE Access the paper "Multi-cell Scheduling in ODMA-based Small Cells" [69] has been published. Also, the papers "Opportunistic Hybrid ARQ – Enabler of Cloud-RAN over Non-Ideal Backhaul" [70] and "Backhaul-aware Energy Efficient Heterogeneous Networks with Dual Connectivity" [71] have been accepted in IEEE Wireless Communication Letters and Springer Telecommunication Systems, respectively. At EuCNC 2014 in Bologna the joint WP3 / WP2 paper "Towards a Flexible Functional Split for Cloud-RAN Networks" [72] was presented. In addition, the papers "Energy Saving Enhancement for LTE-Advanced Heterogeneous Networks with Dual Connectivity" [73], "The Role of Computational Outage in Dense Cloud-Based Centralized Radio Access Network" [74] and "Robust Proportional Fair Scheduling with Imperfect CSI and Fixed Outage Probability" [75] have been accepted at IEEE VTC Fall 2014, IEEE GLOBECOM 2014, and IEEE PIMRC 2014.

3 Functional Split and veNB

3.1 iJOIN Architecture

This section aims to describe the interactions between CTs related to WP3 as well as the interaction of WP3 with WP4 and WP2. Accordingly, we have refined the analysis provided in [5].

The WP3 CTs are listed in Table 3-1 and are classified accordingly to their specific functionalities and functional split configurations. In particular, CTs 3.2 and 3.3 can be characterized as SON functionalities, which enable centralized connection control, and they adapt the system parameters to changes in the cellular network, due i.e. to the network load, energy constraints, and mobility.

The other CTs are used in the centralized resource allocation framework: in particular, CT 3.1 enables optimized BH resource allocation; CTs 3.4, 3.5, and 3.9 are devoted to enhance the performance of downlink transmissions by increasing spectral efficiency, mitigating inter-cell interference, and coordinated RRM. CTs 3.7 and 3.8 increase the robustness of uplink transmissions by using inter-cell cooperation and exploiting spatial diversity. CT3.6 is devoted to investigate the iJOIN Utilization Efficiency metric, which enables to assess the improvements of the proposed CTs. Hence, this classification it is not applicable to CT3.6.

Table 3-1: iJOIN RRM/MAC Candidate Technologies (CTs)

CT	Topic	Abbreviation	Function	Functional Split
3.1	Backhaul Link Scheduling and QoS-aware Flow Forwarding	BH Manager	BH RRM	Centralized Resource Allocation
3.2	Partly decentralized mechanisms for joint RAN and backhaul optimization in dense small cell deployments	Coordinated Cell Selection	SON	Centralized Connection Control
3.3	Energy-Efficient MAC/RRM at Access and Backhaul	EE RRM	SON	Centralized Connection Control
3.4	Computational Complexity and Semi-Deterministic Scheduling	SD Scheduler	DOWNLINK RRM	Centralized Resource Allocation
3.5	Cooperative RRM for Inter-Cell Interference Coordination in RANaaS	Coop. RRM	DOWNLINK RRM	Centralized Resource Allocation
3.6	Assess and Increase Utilization and Energy Efficiency	n/a	n/a	
3.7	Radio Resource Management for Scalable Multi-Point Turbo Detection	MPTD RRM	UPLINK RRM	Centralized Resource Allocation
3.8	Radio Resource Management for In-Network-Processing	INP RRM	UPLINK RRM	Centralized Resource Allocation
3.9	Hybrid local-cloud-based user scheduling for interference control	HL Scheduler	DOWNLINK RRM	Centralized Resource Allocation

Figure 3-1 represents the functional interactions of the WP3 CTs (the blue box) as well as the exchange of information required between WP4 (in red) and WP2 (in green). From WP2, we take into account input and output information from the two main blocks, namely RAN-PHY Functions and BH-PHY Functions. WP3 provides to WP2 RRM and MAC information concerning the radio access and the backhaul, like scheduling maps and link adaptation parameters; WP2 forwards to WP3 estimated radio and backhaul channel information such as SNR and user data after detection and decoding.

The exchange of information between WP3 and WP4 can be divided across two iJOIN logical entities: the iNC and the iTN. WP4 provides to WP3 information about the backhaul configuration and measurements such as routing information and mobility information.

In addition to the two main WP3 blocks discussed above, we identified basic functions that include standard functionalities for the BH and RAN management, which support the iJOIN RRM/MAC enablers.

Finally, we can identify in Figure 3-1 also the interaction of WP3 CTs with the iveC, which, according to the iJOIN architecture [15], is the logical entity that adapts the functional split configuration according to system objectives and constraints.

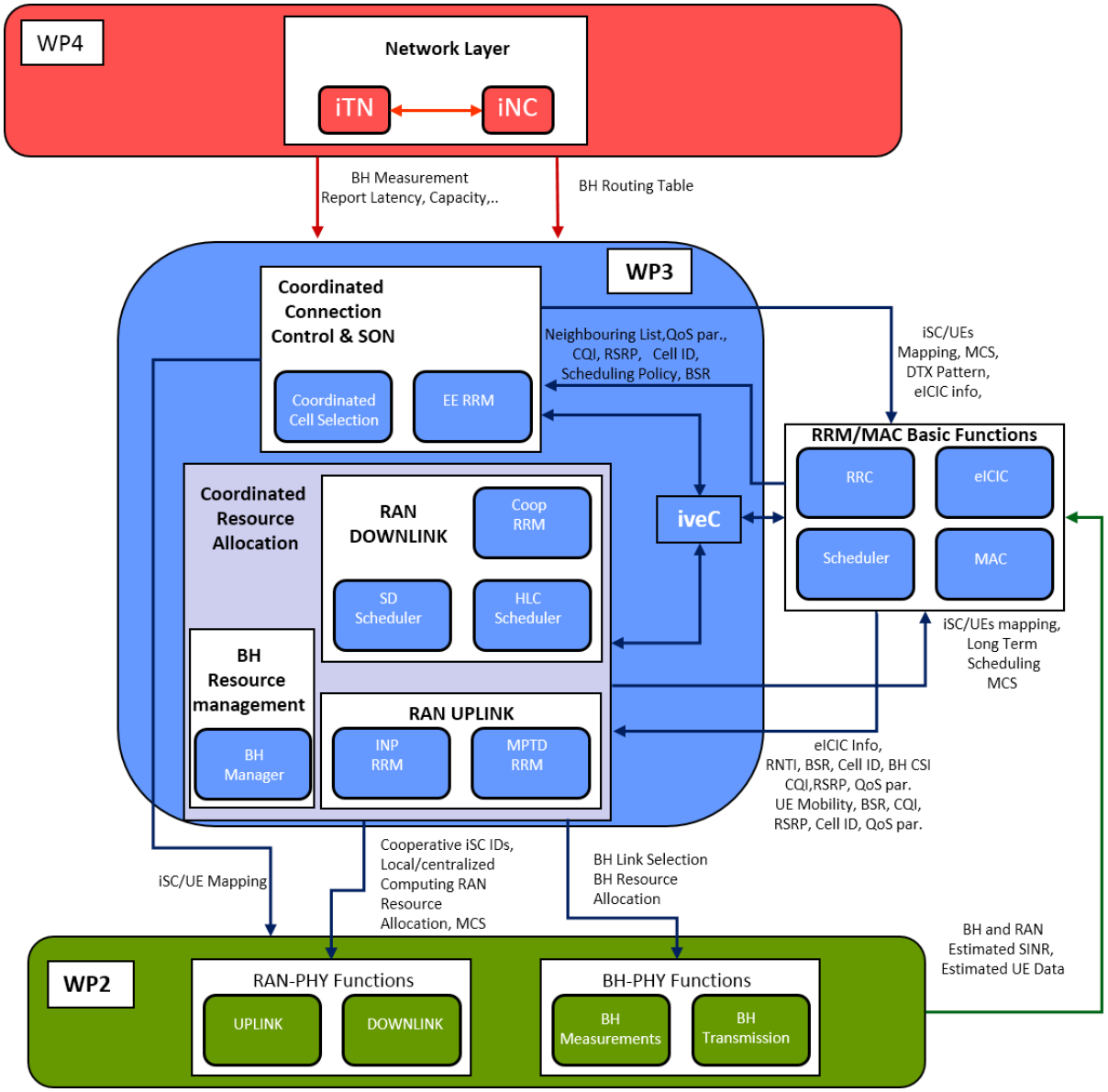


Figure 3-1: WP3 functional architecture.

3.2 Functional Split

Functional split can be seen as a way to exploit the benefits of iJOIN architecture, by adaptively setting and adjusting the level of centralization of key functions at the cloud. To accomplish that target, functional split should have two main characteristics:

- Flexibility: At the veNB, the L1-2 functions can be either located at the RANaaS or at the iSC. The centralization or de-centralization of some of these functions should be decided upon optimizing the network performance subject to some feasibility constraints. These feasibility constraints are the factors that allow us to decide where to perform certain functions and are further discussed below.
- Tune-ability: Another key characteristic we should consider is how tune-able is the operation of specific functions in time domain, so as to capture potential changes at the backhaul availability, traffic load and other factors. In other words, in addition to the flexibility of the functional split (which can be seen as a static optimization problem), we should also consider how the functional split selection problem can be interpreted in a dynamic environment.

To achieve a flexible and tune-able functional split, as a next step, we will provide key decision factors which identify the optimal selection of the functional split taking into account different network limitations. To this end, the decision factors, as well as two key configuration examples, are presented in more detail in the following sub-sections.

3.2.1 Functional Split Decision Factors

The flexibility and tune-ability of the functional split rely on certain factors which can influence the decisions for the level of centralization at the RANaaS. One factor that can strongly influence the functional split decision are the backhaul technology and topology limitations, due to the fact that these can dictate specific functional split options for certain functions. Additionally, other factors can be associated with constraints in order to comply with the 3GPP standardization, minimum centralization requirements imposed by the WP3 CTs and some CT dependencies corresponding to lower layers. Following, this section covers the variables which have a key impact on the functional split selection for L2 functions.

3.2.1.1 Backhaul constraints

In order to decide the functional split configuration, one of the key factors that must be considered is the specific technology used in the backhaul. Depending on the technology, the values of latency and bandwidth in the backhaul will vary, affecting the level of centralization that can be achieved at the RANaaS. In the following, the main technologies that can be used in the backhaul are classified and summarized:

1) Wired Backhaul

Wired backhaul relies mostly on two physical mediums: copper and optical fibre. Considering the copper-based solutions, leased T1/E1 copper lines are extensively used in cellular systems as they can provide suitable support for voice traffic, with deterministic QoS, low latency and jitter. Additionally, xDSL solutions (such as VDSL) can be used when the distance from the small cell is small. However, copper lines do not scale easily to provide adequate bandwidth at distances exceeding few hundred meters to support emerging broadband technologies [26]. On the other hand, optical fibre can provide a multi-Gbps throughput connectivity that can be achieved using point-to-point (PtP), ring/mesh, and point-to-multipoint (PmP) (i.e. gigabit passive optical network) technologies [27]. In general, PtP links and ring/mesh solutions using WDM and Statistical Packet Multiplexing will have better performance than PmP in terms of latency and throughput. Optical fibres are usually deployed in urban and sub-urban areas where very high traffic-carrying capacity is required. Although a fibre-based backhaul offers long-term support with respect to increasing capacity requirements, this comes at a relatively high cost.

2) Wireless Backhaul

Various wireless backhaul solutions exist with diverse characteristics in terms of the type of propagation, the spectrum used and the network topology. In general, the advantage of wireless backhaul is the freedom from cabling, which is expensive to deploy due to the high costs of installation. Wireless solutions need only equipment at the small cell and the Point of Presence[1] (PoP) offering reduced costs and faster deployment. The main categories of wireless technologies are the following:

- Sub-6 GHz: This category can be seen as a '*Non Line of Sight'* (NLoS) category and includes carrier frequencies below 6 GHz (3.5 GHz licensed and 2.4 / 5.8 GHz unlicensed). Sub-6 GHz backhaul can be easy to plan and deploy in urban areas, thereby significantly reducing the cost and duration of small cell network roll out. In particular, the 3.5 GHz band has emerged as a promising candidate for the dedicated use of small cells. On the other hand, the unlicensed spectrum provides a large amount of freely available bandwidth but is likely to be already (or later) heavily used by Wi-Fi hotspots, Bluetooth and other equipment.
- Free-space optical (FSO): FSO backhaul is a '*Line of Sight'* (LoS) technology that uses invisible beams of light to provide optical bandwidth connections at multi-Gbps rates [28]. FSO uses the same transmission wavelengths as fibre optics (850 nm, 1550 nm) but transmits over the air. Its fundamental similarities to fibre optic make it a strong candidate to support future packet-centric

[1] Points of Presence are defined as logic entities which offer connectivity to the core network for small-cells.

networks. However, its main drawback is the requirement of high-stability mounting and high path-loss due to obstructions and fog attenuation.

- Microwave Backhaul: Microwave radio can be seen as an alternative choice of backhaul connectivity especially in areas where a wired connection is not available. Microwave transmission operates mainly in licensed spectrum (28 GHz to 42 GHz) and requires LoS (or near-LoS) [26]. In general, microwave radio can provide capacity of some hundred Mbps [29] and high availability especially in higher bands.
- Millimetre wave (mmW) radio: Conceptually, mmW-radio refers to any RF technology operation in the 30-300 GHz range, but it is generally used to discuss 60-80 GHz, also known as "E-band" [30]. In this context, several GHz-wide bandwidths are available and can provide multiple Gbps even with low-order modulation schemes. In addition to these high-data rates, mmW radio band can offer excellent immunity to interference, high security and the reuse of frequency. mmW radio requires clear LoS propagation and its range is restricted by the oxygen absorption which strongly attenuates ≥60 GHz signals over distance. Therefore, high gain directional antennas are used in order to compensate for the large free space propagation losses.

There are two main topology types applicable to most of the wireless backhaul technologies: 1) *Point-to-Point (PtP)* and; 2) *Point-to-Multipoint (PmP)*. In PtP, individual point to point links between nodes (i.e. access points or gateways) can be interconnected to form chain, tree, ring, or mesh topologies, whereas in PmP a PoP forms multiple links to a number of access points. The main challenges of PtP are: a) the large number of antennas that may be required at the PoPs; b) the requirement for frequent re-planning when new nodes are added; c) the inclusion of redundant links offering resiliency to link outages and; d) multi-hop links can lead to latency restricted performance. On the other hand, PmP links may be more efficient to pool resources across a larger, changing number of nodes and average out any difference in traffic demand at different times of day. While PtP topologies can be used in all the technologies listed above, PmP is used only in Sub-6 GHz and Microwave.

Table 3-2 summarizes some key features of the discussed candidate backhaul technologies, based on the classification described in [21]. Regarding the latency classification, the second column refers to the RANaaS-to-iSC latency, while the third one to the per-hop latency. The data for this per-hop latency is taken from D4.2 [59], where full information regarding per-hop parameters for different backhaul technologies can be found.

Since the latency data in D4.2 is more exhaustive, the following assumptions have been performed:

1. For ideal fiber access, DSL access and Cable, the per-hop latency is equal to the total latency. Therefore, we are assuming that there are no intermediate nodes between iSCs and the RANaaS entity.
2. Fiber Access 1 and 2 are considered to use technologies based on Metro Optical Network or Passive Optical Networks with several intermediate nodes. The resulting latency employing these technologies is mainly due to processing at the optical nodes.
3. In all the wireless technologies we assume PtP communication and FDD multiplexing for mmW radio and microwave. We consider that there can be some intermediate nodes when wireless transmission is employed in the backhaul. A more detailed presentation of the backhaul classification is also included in Appendix III.

Table 3-2: Backhaul Classification (based on [21] and D4.2 [59])

BH technology	Total Latency (one-way)	Per-hop Latency	Throughput
Ideal fiber access	2.5 µs	5 µs/km	10 Gbps
Fiber Access 1	10 – 30 ms	≤1 ms	10 Mbps – 10 Gbps
Fiber Access 2	5 – 10 ms	≤1 ms	100 Mbps – 1 Gbps
DSL Access	5 – 35 ms	5 – 35 ms	10 Mbps – 100Mbps
Cable	25 – 35 ms	25 – 35 ms	10 Mbps – 100Mbps
Sub-6 GHz Wireless	5 – 10 ms	≤5 ms	50 Mbps – 1Gbps
Microwave	< 1 ms	≤200 µsec	100 Mbps – 1Gbps
mmW radio	< 1 ms	≤200 µsec	500 Mbps – 2Gbps

3.2.1.2 3GPP requirements

1) Bandwidth Requirements

The bandwidth required for backhauling between an iSC and the cloud generally depends on a large number of parameters, such as the number of sectors, the number of carriers, the bandwidth of the carriers and the load of iSCs. In addition, it depends on the functional split itself [31].

2) Latency Requirements

If the backhaul can fulfil the bandwidth requirements for a given functional split, it will also inevitably add some latency. Since 3GPP defines many timers from the MAC to the RRC layer, these values will ultimately define the maximum latency requirement needed per layer enabling a transparent functional split, i.e. without any specification changes. Table 3-3 shows the main specified timers and timing constraints per OSI layer. Grey rows have been identified as timers not being "relevant" to the functional split, i.e. (either they do not run at the E-UTRAN side or they have a long timing range definition). In general, the higher layer we go, the higher timing range and time we have. In practice, if the latency needed for the MAC layer is fulfilled by the backhaul, then any functional split at the MAC layer or above will be possible.

The LTE MAC layer defines how much data is taken from RLC queues into MAC transport blocks, depending on channel conditions and available resources. If the latency on the backhaul link is low, there is no significant impact. However, in the case of high backhaul latency, the actual preferred link adaptation and therefore transport block size of the MAC layer may be outdated at the point in time when RLC prepares the PDU for the MAC layer. This can lead to an increased outage and re-transmissions due to imperfect link adaptation. A more conservative choice of MCS may solve the problem but at the cost of a lower throughput, which is addressed by CT3.4 in Section 4.4.

More importantly, the LTE MAC layer has strict HARQ timing, especially in the uplink. Indeed, the scheduling of an uplink transmission at a subframe *n* is done at subframe *n*-4. Once a packet has been sent at subframe *n* for a given HARQ process, an acknowledgement (positive or negative) is expected at subframe *n*+4 such that the HARQ process can either be used at subframe n+8 for a new transmission or a retransmission (hence the 8ms of HARQ RTT in Table 3-3). Due to the synchronous nature of HARQ in the uplink, any functional split at the base station MAC layer requires the round-trip time plus the processing to be done in below 3 ms, which is a tight constraint. This is illustrated in Figure 3-2 with an example for a FDD system with a symmetric one-way backhaul latency of just under 1 ms, resulting in a processing delay budget at the RANaaS of slightly more than 1ms. Note that the air interface delay is equivalent to the propagation delay, which is constrained by the maximum timing advance value in LTE of 532.48 µs. In this example, the air interface delay is therefore too high. Nevertheless, the range of acceptable backhaul delay values and also the RANaaS processing delay budget is strongly constrained. Furthermore, in this example it is assumed that UL/DL subframes are time-synchronized at the iSC.

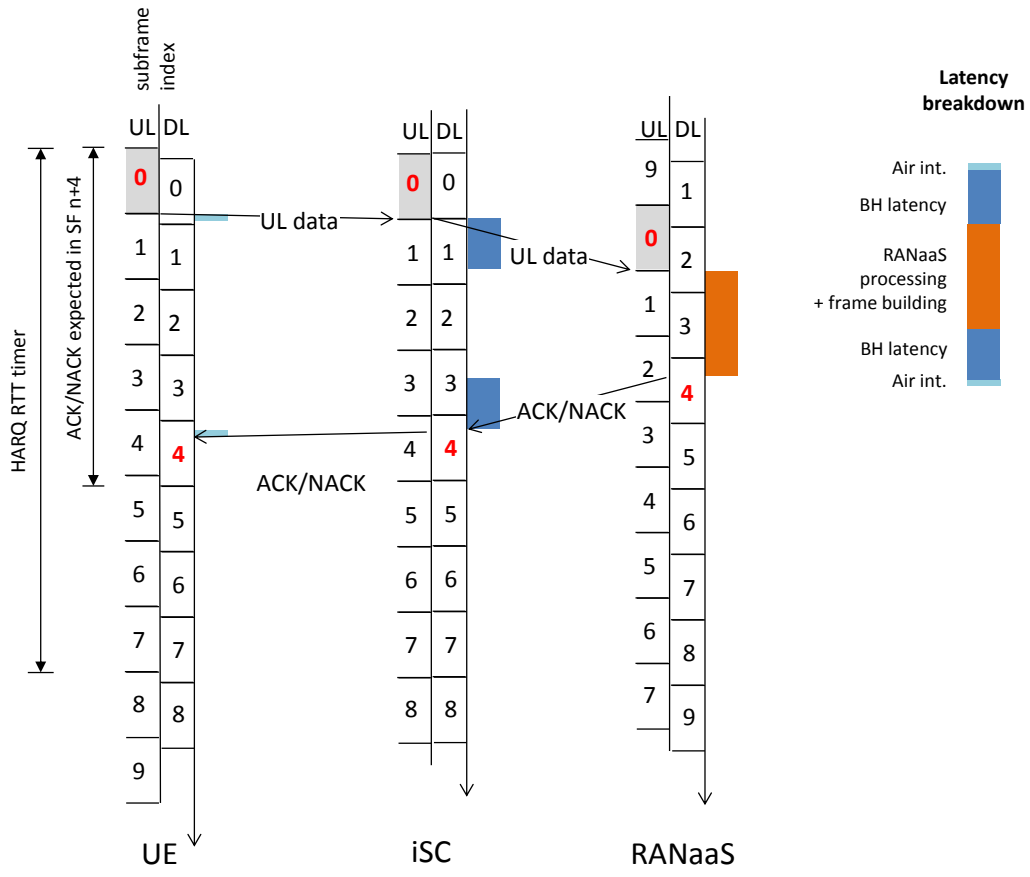


Figure 3-2: Break down of timing constraints for centralized HARQ in LTE FDD with < 1 ms one-way backhaul latency and ~ 2ms RANaaS processing and frame building delay.

Backward compatible solution exists to delay the retransmission but this will stall the HARQ process at the same time, reducing the throughput [31]. Thus, the uplink HARQ timing constraint appears to be the most critical one for any functional split at or below the MAC layer, if a compliant LTE-solution is needed with no performance degradation. . One way to break the HARQ constraint is to apply opportunistic HARQ as described in [68].

Table 3-3: 3GPP timing requirements

	Timer	Purpose	Min	Max	Default	Relevant to Functional Split
MAC [44]	HARQ UL indication	When an ACK/NACK indication is expected	In FDD: SF+4 (3 ms) In TDD: depends on configuration, maximum SF+7 (6 ms)			**Yes**
	HARQ RTT Timer	When an HARQ process is available	In FDD: 8ms In TDD: k+4ms, where maximum k = 6ms			**Yes**
RLC [45]	t-PollRetransmit	For AM RLC, poll for retransmission @tx side (if no status report received)	5ms	500ms	45ms	**Yes**
	t-Reordering	For UM/AM RLC, RLC PDU loss detection @rx side	0ms	200ms	35ms	**Yes**

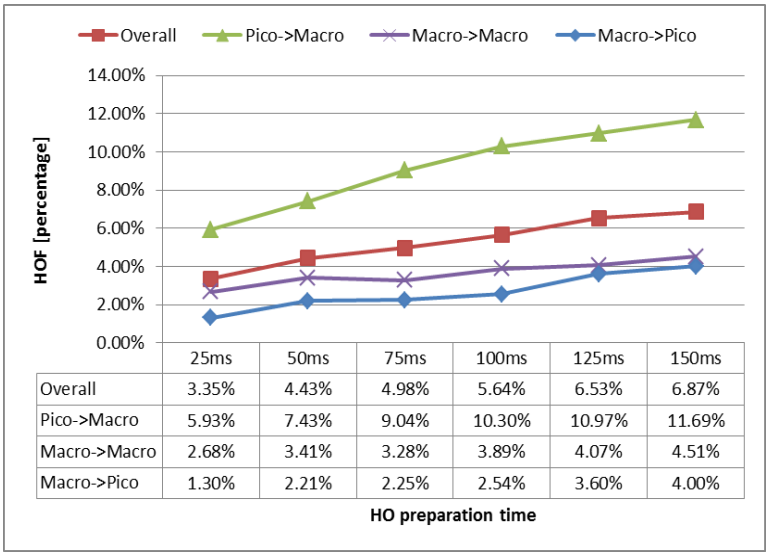
	t-StatusProhibit	Prohibit generation of a status report @rx side	0ms	500ms	0ms	**Yes**
PDCP [46]	discardTimer	At UE side in UL. Start at reception of PDCP SDU from upper layer. Discard PDCP SDU / PDU if expiration or successful transmission	50ms	Infinity		**Yes**
RRC [47]	T300	RRCConnectionRequest. If expire, reset MAC & signal RRC connection failure	100ms	2000ms		**Yes**
	T301	RRCConnectionReestablishmentRequest If expire, go to RRC_IDLE	100ms	2000ms		**Yes**
	T302 ac-BarringTime	RRCConnectionReject If expire, inform upper layers about barring alleviation	(0.7+ 0.6 * rand) * ac-BarringTime 4s	 512s		**No**
	T303	Access barred (mobile originating calls) If expire, inform upper layers about barring alleviation	Same as T302			**No**
	T304	RRCConnectionReconfiguration Cell change order in MobilityControlInfo Cell change order in MobilityFromEUTRACommand	50ms 100ms	2000ms 8000ms		**Yes**
	T305	Access barred (mobile originating signalling) If expire, inform upper layers about barring alleviation	Same as T302			**No**
	T306	Access barred (mobile originating CS fallback) If expire, inform upper layers about barring alleviation	Same as T302			**No**
	T310	Detection of physical problem (successive out-of-sync from lower layers) If expire, if security not activated, go to RRC_IDLE, else initiate connection reestablishment	0ms	2000ms	1000ms	**Yes**
	T311	RRC connection reestablishment (E-UTRA or another RAT). If expire, go to RRC_IDLE	1000ms	30000ms	1000ms	**Yes**

	T320	RRCConnectionRelease UE to use dedicated cell reselection priority parameters. If expire, discard dedicated parameters & use broadcasted parameters	5min	180min		**No**
	T321	Duration during which the UE is requested to perform measurement measConfig including a reportConfig with the purpose set to reportCGI If expire, send measurement reports	150ms if for E-UTRA handover 1s if for E-UTRA 2s if for UTRA FDD handover 1s if for UTRA TDD handover 8s if for UTRA 8s otherwise			**No**
	T330	Duration during which the UE is requested to perform measurement logging	10min	120min		**No**

3) Protocol Requirements

The first function of interest is cell (re)selection. Cell (re)selection is located in the RRC layer and belongs to the control plane protocol stack. This process allows selecting for each UE the best cell that can serve it. For this purpose, each UE measures the received signal strength of the different surrounding cells. Based on these values, the UE RRC (in idle mode) or the BS RRC (in connected mode) will select the strongest one from the list and will initiate the cell (re)selection/handover procedure.

The main challenge of the cell (re)selection process is that the current associated mechanisms are based solely on the power level received from neighbouring cells, without using information regarding the cell loads and backhaul capacities. Figure 3-3 illustrates an example of the impact of backhaul latency on the performance of cell (re)selection in the case of handover due to mobility. It can be observed that a higher backhaul latency increases handover preparation time, which leads to an increasing handover failure rate, especially in case of pico-to-macro cell handovers. This is especially a challenge in dense networks due to the increased handover rate. While corresponding timer values can be adjusted individually for each deployed small cell, this approach does not seem suitable for large deployment scenarios.



	25ms	50ms	75ms	100ms	125ms	150ms
Overall	3.35%	4.43%	4.98%	5.64%	6.53%	6.87%
Pico->Macro	5.93%	7.43%	9.04%	10.30%	10.97%	11.69%
Macro->Macro	2.68%	3.41%	3.28%	3.89%	4.07%	4.51%
Macro->Pico	1.30%	2.21%	2.25%	2.54%	3.60%	4.00%

Figure 3-3: Impact of latency on hand-over failure rate (HOF)

Segmentation and reassembly are also functions of interest to be considered. Both are located in the RLC layer at the transmitter. On the other hand, only the reassembly function is located in the RLC layer at the receiver. The RLC layer is, together with the PDCP layer, responsible for the link reliability functionality such as re-transmissions and re-ordering. The first challenge is the backhaul reliability and its impact on the

3GPP LTE performance. One possibility is to handle errors using standard mechanisms on the RLC layer even though this implies unnecessary overhead on the wireless interface between user terminal and base station. An alternative solution is to re-transmit on the backhaul in order to reduce both delay and overhead on the wireless link. The second challenge is jitter on the backhaul link which adds up to the end-to-end jitter. Hence, the timers maintained by the base station may need to be adjusted in order to compensate the increased jitter. In particular, the base station needs an interface to the network controller in order to receive an estimate of the jitter on the backhaul link.

3.2.1.3 Minimal centralization requirements of CTs

One of the key decision variables for functional split can be the centralization requirements for each CT. In WP3, no centralized processing of user plane data is necessary. Two main configurations are common to all CTs: coordinated resource allocation (CRA) and centralized connection control (CCC). As can be seen in Table 3-4, most of the CTs are associated with CRA, whereas CT3.2 and CT3.3 can be associated with CCC.

Table 3-4: Mapping of configurations to CTs

	Coordinated Resource Allocation	Centralized Connection Control
CT 3.1	☑	
CT 3.2		☑
CT 3.3		☑
CT 3.4	☑	
CT 3.5	☑	
CT 3.6	N/A	N/A
CT 3.7	☑	
CT 3.8	☑	
CT 3.9	☑	

Coordinated Resource Allocation: In this scenario, Inter-cell RRM for UL/DL is performed in RANaaS. For the downlink case CQI feedback is sent by each iSC over J1, informing RANaaS about all the users' channel conditions. Thereafter, RANaaS processes all this information and dynamically allocates RBs to iSCs and users. On the other hand, for the uplink case, the overload indicator, interference, and power measurements for each RB would be sent to RANaaS in J1 by each iSC. Following, RANaaS will send the resource allocation decisions in J1 to all the corresponding iSCs.

In Figure 3-4, the mapping of this functional split scenario to the protocol stack of the involved entities is illustrated. These entities are the RANaaS, the iSC and iTN which can be optionally co-located with the iSC to serve the small cell backhaul. We observe that most of L2 functions are performed in the iSC, whereas the resource coordination is performed at the RANaaS entity. Here, it should be mentioned that the resource coordination function encapsulates both the resource allocation for the access and the scheduling of the small cell wireless backhaul links.

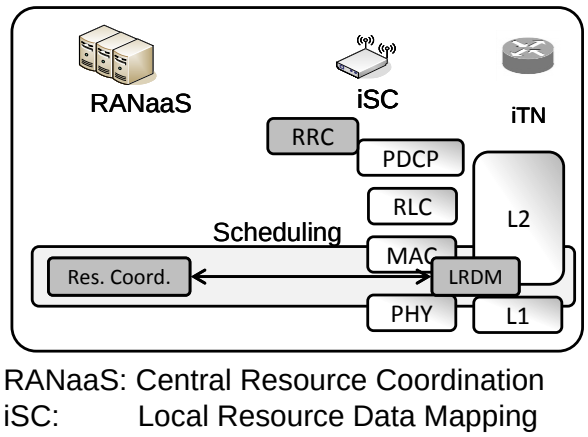


Figure 3-4: Mapping of the functional split scenario to the protocol stack for CRA

Centralized Connection Control: In this scenario, the handover process is fully performed in the RANaaS. The serving iSC sends the UE Measurement Report to RANaaS through J1, which performs the handover decision based on this report and using the load information and backhaul restrictions of the neighbouring cells. This decision is sent to the serving iSC and the new iSC through J1. Additionally, the Handover Request to the new iSC and the admission control can be avoided since RANaaS also takes care of the admission control of the new iSC. The rest of the process corresponds to an intra-eNB handover (only sending of unacknowledged DL packets and out of sequence UL packets may be required through J2 if management of U-plane information is performed locally in the iSC). Finally, more complex solutions can be exploited if the reallocation of users of the new iSC is allowed.

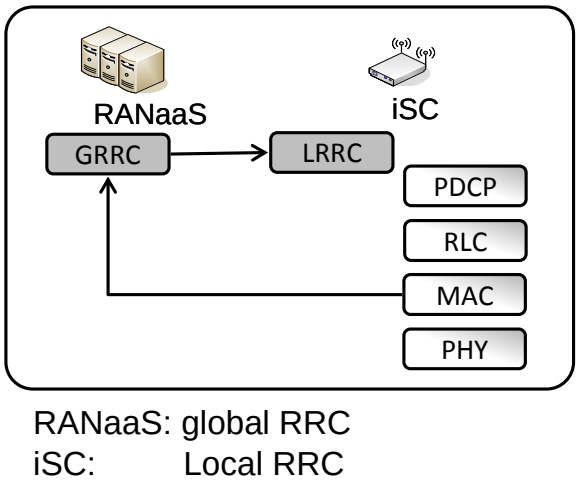


Figure 3-5: Illustration of Centralized Connection Control

Figure 3-5 illustrates the functional split for CCC which is mainly used by CT3.2 and CT3.3. Here, the RRC can have two variants: the Global RRC (GRRC) and the Local RRC (LRRC). In this configuration, the Global RRC operates in RANaaS and performs some basic functions, i.e. the Bearer setup and the UE attachment. On the other hand, LRRC is handled at the iSC and is responsible mainly for acquiring measurements, handover decisions, etc. More details about LRRC functionalities can be found at the CT3.2 description.

3.2.1.4 Lower layer dependencies

The flexible functional split also allows the centralization of PHY layer processing, as proposed in D2.1 [14].

In the downlink, multi-cell precoding (comprising users of multiple iSCs) can be performed at the RANaaS in order to mitigate inter-cell interference. Such techniques require precise knowledge of the instantaneous channel fading state, which need to be exchanged via J1 links. Due to latency on the backhaul, channel state information (CSI) becomes outdated and the precoder is only imperfectly aligned to the actual channel, resulting in performance losses. Consequently, the gains obtained by centralized precoding need to be compared with the performance degradation due to backhaul latency.

In addition to latency, a limitation in the capacity of the backhaul also affects the performance obtained from PHY layer centralization. This basically comprises two aspects. Firstly, the transmission of CSI from the iSCs to the RANaaS. Additional compression might be required, leading to a further reduction of the CSI. Secondly, the transmission of the pre-coded data (or alternatively the precoding matrix itself) from the RANaaS to the respective iSC. In this case, quantization can act as a further source of impairment.

Performing centralized multi-cell processing in the uplink, the receive filter as well as the transmit power allocation can be performed at the RANaaS. The multi-user reception is not directly affected by backhaul latency since channel measurements and data can be extracted from the same RBs. In contrast to the downlink, the power control computation suffers from outdated channel information, which results from feedback and backhaul latency.

In general, performing PHY layer processing at the RANaaS affects also higher layer functions. In particular, scheduling and RRC should be performed at the RANaaS in order to avoid additional latency and J1 traffic.

3.2.2 Functional Split Options

Figure 3-6 illustrates different functional split options of the LTE protocol stack including MAC and layers above. Split options C.1 and C.2 enable coordinated resource allocation by centralizing parts (option C.1) or the full MAC scheduler (option 2) into the RANaaS entity. Split options D.1 and D.2 enable coordinated connection control by centralizing at least radio resource control (option D.1) or additionally the PDCP layer (option D.2). Note that split options D correspond to the split bearer options defined for the dual connectivity feature defined in LTE Rel. 12. This feature is illustrated Figure 3-7, which contains different options for centralizing (at a Macro/Master eNB – MeNB) parts of the LTE protocol stack, including RRC and PDCP[2]. In the following, we describe each split option in more detail and summarise their pros and cons.

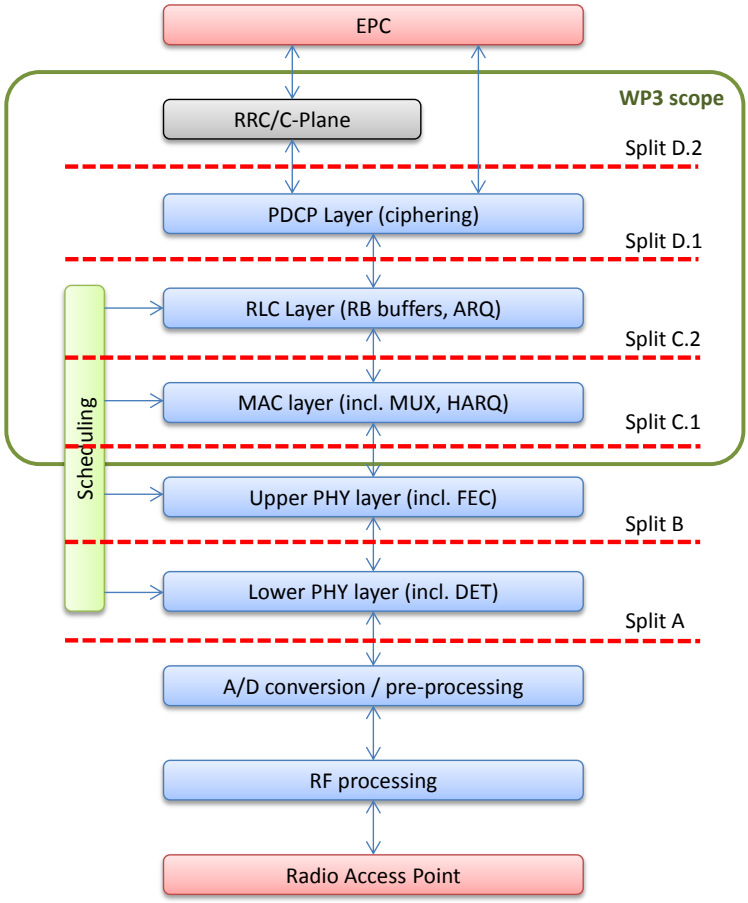


Figure 3-6: Functional split options on MAC layer

[2] Note that at time of writing, dual connectivity was not yet officially adopted in 3GPP technical specifications. However, the feature is technically endorsed and will be part of LTE Rel. 12.

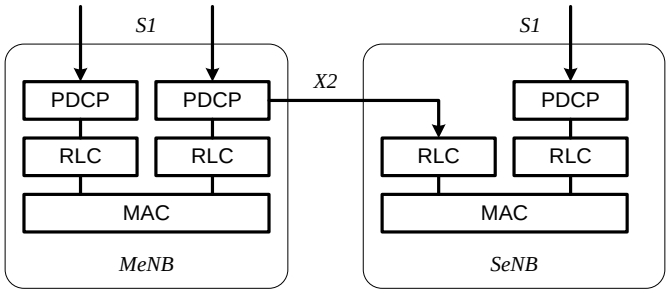


Figure 3-7: Dual connectivity as developed by 3GPP [61]

1) Split option D.2 is characterized by a centralized RRC layer. This allows for centralized connection control and SON-like functions such as parameter tuning (e.g. for hand-overs, power control, etc), load balancing as proposed in CT 3.2 and efficient power management as proposed in CT 3.3. Note that this is the only option where only the C-Plane is centralized, although the U-plane terminating points are in practice often co-located with the corresponding C-plane interfaces. From an interface and architectural point of view this option is equivalent of implementing the X2-C interface as defined in LTE Rel. 12 (see Figure 3-8), assuming that the MeNB is located at the RANaaS platform. The backhaul requirements of this split option are constrained by RRC timers, which are configurable in the range of a few hundred milliseconds to seconds (c.f. Table 3-3).

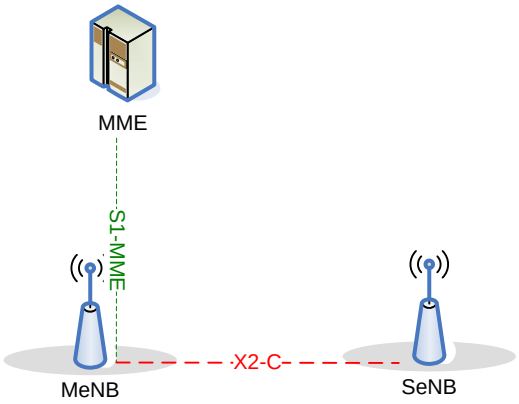


Figure 3-8: X2-C interface [61]

- **Pros:** Small functional impact, compliant to LTE Rel. 12, low backhaul requirements, no buffering of user-plane data at RANaaS required.
- **Cons:** limited centralization gains, mobility visible to Core Network (CN) in case of inter-iSC HOs, ciphering on PDCP layer in both RANaaS and iSC.

2) Split Option D.1 centralizes additionally to Option D.2 also the PDCP protocol layer. This corresponds to the "split bearer" option of the dual connectivity feature. It has similarly as Option D.2 only a small impact on the LTE protocol stack, although it requires potentially buffering of PDCP packets at the RANaaS, thus introducing a two-stage queuing system since also RLC maintains buffers for each radio bearer. Backhaul requirements are constrained by the PDCP discard timer, which has a minimum configurable value of 50ms.

- **Pros**: small functional impact, compliant to LTE Rel. 12, low backhaul requirements, ciphering in RANaaS, inter-iSC mobility not visible at CN.
- **Cons:** limited centralization gains, PDCP buffering at RANaaS.

3) Split Option C.2 enables coordinated resource allocation by centralizing RLC and parts of the radio resource allocation. Specifically, Hybrid ARQ is located at the iSC, while RLC buffering and some more coarse-granular, long-term resource allocation is located in the RANaaS. This "split-scheduling" approach has therefore less stringent requirements on backhaul latency, while enabling centralization gains by means of interference mitigation techniques. However, it requires a new mechanism for resource allocation preferably on the granularity of a few frames, to avoid strong latency dependencies on the one hand and out-dated channel information on the other hand. Dedicated signalling for resource allocation commands is

required. Additionally, this approach may require buffering of MAC PDUs at iSC level. This approach has therefore a higher impact on the overall LTE protocol architecture compared to other split options.

- **Pros**: enables high centralization gains, still low backhaul latency requirements, inter-iSC mobility not visible at CN
- **Cons:** requires dedicated signalling, buffering at MAC layer

4) Split Option C.1 enables like Option C.2 coordinated resource allocation, but the MAC layer including HARQ is now fully centralized at the RANaaS platform. This means that resource allocation on resource block (RB) level is performed at the RANaaS entity. This approach corresponds to the Femto Application Platform Interface (FAPI) architecture defined by SmallCell Forum [62]. The strong requirements of HARQ on latency of 3ms apply (see Section 3.2.1.2) and additionally any delay jitter must not exceed the configured interval of PDCCH DCI information in the subframes.

- **Pros:** enables high centralization gains, inter-iSC mobility not visible at CN, no buffering at iSC, compliant to FAPI approach, no dedicated signalling required
- **Cons:** strong latency requirements

Table 3-5 provides a comparison of the different split options together with an overview of backhaul requirements, functional impact and expected centralization gains. The following observations can be made.

With the functional split getting lower in the protocol stack:

- requirements on backhaul latency increase;
- requirements on bandwidth are unaffected (in WP3 scope);
- number of applicable CTs increase;
- centralization gains increase.

The impact on LTE depends on the individual split and cannot be directly correlated with protocol level.

It can be concluded that with a low-latency backhaul, split option C.1 is preferable due to the low impact on the LTE protocol stack and the full enablement of centralization gains. If backhaul RTT values are above 3ms, split options C.2 and D.1 are preferable, depending on the optimization goals of the deployment scenario. Option D.2 is not preferable in the context of iJOIN RANaaS scenarios due to the additional burden of ciphering in the iSCs, and the potential visibility of inter-iSC handovers to the CN.

Table 3-5: Comparison of split options

Split option	Lowest layer centralized	Impact on LTE	RTT requirements	Bandwidth requirements	Centralization scheme	Applicable CTs	Main centralization gains
D.2	RRC	Small, ciphering in iSC	Several hundred milliseconds to seconds	U-plane + C-plane overhead	Centralized connection control	CT 3.2, CT 3.3	Load balancing, energy efficiency
D.1	PDCP	Small	> 50ms	U-Plane + C-Plane overhead	Centralized connection control	CT 3.2, CT 3.3	Load balancing, energy efficiency
C.2	RLC + long-term scheduling	Split scheduling, dedicated signalling for resource allocation	Several frames (10ms each)	U-Plane + C-Plane overhead	Coordinated resource allocation	CT 3.2, CT 3.3, CT 3.4, CT 3.7, CT 3.8, CT 3.9	Interference mitigation, cooperative schemes
C.1	MAC	Small	<3 ms (HARQ)	U-Plane + C-Plane overhead	Coordinated resource allocation	all	Interference mitigation, cooperative schemes

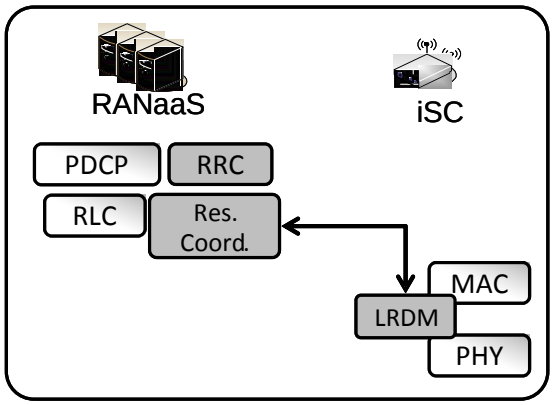
3.2.3 Functional Split Configurations

Functional split can be realized at the MAC layer to enable coordinated RRM and centralized scheduling [5], which is closely coupled with dynamic inter-cell interference management and specifically to interference coordination. ICIC has the task to manage radio resources such that inter-cell interference is kept under control. ICIC is inherently a multi-cell RRM function that needs to take into account the resource usage status and traffic load situation of multiple cells. The preferred ICIC method may be different in the uplink and downlink. This approach increases the overall system spectral efficiency by mitigating inter-cell interference and exploiting multi-user diversity.

The backhaul capacity requirements associated with a full centralized RRM approach is still high, since sharing channel state information is necessary to correctly implement, i.e., a multi-cell scheduler. Moreover, performance depends also on the backhaul latency, since outdated channel state information (CSI) strongly limits the achievable gains.

In LTE, dynamic inter-cell interference management is supported based on messages exchanged between neighbouring cells over the X2 interface. In the context of RANaaS, there are two main options to perform inter-cell RRM in a centralized way. One possibility as described by Functional Split Option C.1 is to perform the resource allocation on RB granularity centrally to minimize inter-cell interference. The actual schedule plan is then forwarded to the corresponding iSCs. The second option is to resolve only inter-cell interference conflicts, i.e. small-cells perform local scheduling and in the case of significant inter-cell interference, a coarse-gain central schedule is performed and ex-changed with the small cells. The performance of the first option is higher but it also imposes stronger requirements on the backhaul latency. In contrast, the second option, which represents a two-stage scheduling approach, copes with higher backhaul latency while preserving a major part of the gains.

The following figure shows an example functional split configuration for CRA, where the backhaul latency can be a limiting factor. In this case, Global Resource Coordination, RRC and PDCP are performed at the RANaaS entity, whereas lower MAC is performed at the iSC, corresponding to functional split option C.2.



CRA: Coordinated Resource Allocation

CCC: Centralized Connection Control

LRDM: Local Resource Data Mapping

Figure 3-9: Example functional split configuration for CRA

Coordinated RRC enables to deal with user mobility, to optimize cell load and to perform cell activation/deactivation mechanisms for energy saving purposes. The PHY/MAC adapting mechanisms are implemented in short-time scale (from milliseconds to seconds) to reply to fast changes due to the channel conditions and traffic; however, coordinated RRC operation is characterized by less stringent constraints in terms of required overhead and timing.

Some centralization approaches would not impose strong latency requirements on the backhaul. In the iJOIN framework, a coordinated load balancing mechanism has been proposed to distribute the cell load amongst neighbouring cells by jointly taking into account the radio access and the backhaul capacity [5]. This approach results in notable throughput improvement, especially in highly loaded scenarios. Moreover, a mechanism to control the cell activity has been introduced to enhance the system energy efficiency.

Neighbouring cells and their backhaul links are switched-on and off, according to the actual cell load and QoS constraints. Note that in both these solutions the RRM and the lower functionalities are locally implemented at each small cell.

Nevertheless, such gains based on centralized connection control can naturally also be implemented if lower layers are centralized if the backhaul constraints allow so. This is illustrated in Figure 3-10, where an example functional split configuration with CCC is used together with PHY centralization.

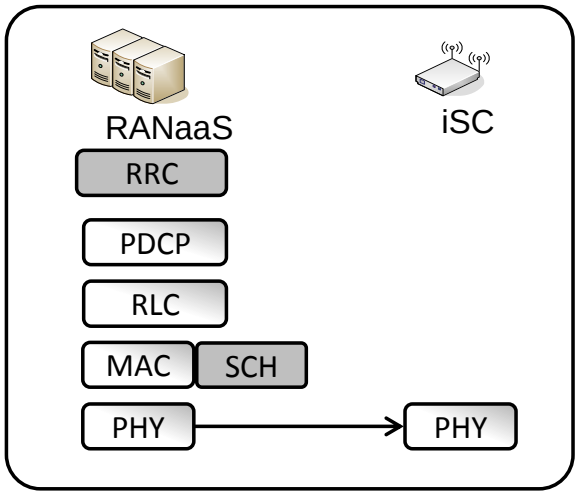


Figure 3-10: Illustration of CCC in case of centralized PHY and MAC processing

3.3 Virtual eNB Implementation

3.3.1 CT Virtualization

Moving a candidate technology (or a legacy 3GPP LTE stack function) from the standard eNodeB to an iJOIN virtual eNodeB means realizing a functional split, as described in the previous section, executing part or the whole CT functionality on the infrastructure where the RANaaS layer is actually hosted and implemented.

In iJOIN, the initial target chosen for the RANaaS layer implementation is a cloud computing datacentre, delivering general purpose computational resources through mechanisms of resource virtualization and sharing, exemplary for the IaaS (Infrastructure as a Service) cloud model. In such model, the cloud computing management layer, based on the actual resource demand from cloud users, instantiates *virtual objects*, essentially computational resource bundles appearing to the respective user as self-standing physical servers, storage units or network resources. *Virtual servers* (a.k.a. *virtual machines*) are pre-configured with basic *software images* including an operating system and additional software modules. The creation of such virtual machines is done by a software component named *hypervisor*, realizing the "translation" from underlying physical resources to virtual resources.

Conceptually, the RANaaS layer on a virtual eNodeB can be implemented on any cloud IaaS platform. In iJOIN, the selected platform for the first RANaaS implementation is OpenStack (www.openstack.org), one of the emerging frameworks in the current cloud landscape. OpenStack offers all the supporting functions needed to create, configure, activate and deactivate the RANaaS virtual objects. The underlying physical infrastructure is made up by standard blade servers, plus storage and network equipment and server virtualization is obtained using Kernel-based Virtual Machine (KVM) hypervisor. This implementation has been chosen for iJOIN, since it allows to run a significant evaluation of the functional split performance in a very mainstream cloud computing environment.

Implementing a candidate technology on the virtual eNodeB model means moving the execution of its functional algorithm (or part of it, in cases where the algorithm can be split into different threads) from the standard eNodeB (or from a legacy small cell) to the cloud platform where the RANaaS layer is running. This has some key implications:

- We are moving a computational workload from an embedded DSP platform into a general purpose computing environment; this porting is not automatic, and demands a recoding of the algorithm to the target platform and to its operating environment;

- The processing flow must get across the backhaul (through the J1 interface), to reach the RANaaS layer from the iJOIN small cells and vice versa; hence, the backhaul perfomance is a key constraint to take into account in moving a candidate technology to the veNB. The iJOIN architecture, based on a SDN-controlled backhaul, supports this need.

The actual feasibility or effectiveness of functional split inside the veNB depends upon key parameters different for the individual candidate technologies because they are tied to the characteristics of each algorithm in terms of distribution, computational intensity and timing. However, generally speaking, the two most significant parameters to consider are *processing power* and *latency*.

Processing power can be a critical parameter for CPU bound algorithms, since general purpose CPUs like the ones powering industry standard servers cannot generally reach the top processing performance rates which a DSP (or even more an ASIC or a FPGA) can achieve. The limitation is both in the CPU own computational power, and in the fact that industry standard servers don't execute microcode but software programs whose interaction with the processor is mediated by an operating system, and are written in non-machine languages which poses a performance penalty. It is true that, on a theoretical standpoint, you could write micro-code programs and bypass the operating system to execute them directly at machine level. However, this wouldn't overcome the possible gap of CPUs versus DSPs, and in the end it would wreak havoc the ultimate sense of functional split. A way to overcome processing bottlenecks in the RANaaS is to redesign (wherever possible) the algorithm to allow parallel execution over more virtual machines. A cloud is theoretically scalable with no limits, hence increasing the grade of parallelization can potentially overcome processing power issues.

Moreover, latency issues can occur at the J1 interface between iSCs and RANaaS, and also among different virtual machines executing a unique candidate technology in distributed mode. The latency is due to the sum of two different elements:

- Transport latency: due to the backhaul in case of the J1 interface, or to the intra-datacentre network in case of virtual machines;
- Software latency: the virtual network mechanisms through which the virtual machines communicate among them and with the outside world could introduce delays in data reception and processing.

Transport latency can be mitigated by improving the backhaul capacity in terms of available bandwidth, and the performance of intra-datacentre networks. Clearly, the former may conflict with one of the key iJOIN objectives, i.e. the ability to optimize transmission quality whatever the backhaul type is.

Software latency is less easy to act on in a standard software and virtualized environment. Basically, improving software latency would require to bypass one of the software layers and going straight to CPU level. In the longer term, server technology evolutions will help to circumvent most latency problems. In the iJOIN scope, one possibility to investigate could be the provisioning of bare metal servers instead of virtual machines. This is one of the emerging paradigms in the cloud computing domain.

Latency could result as a showstopper in particular with those algorithms requiring a tight and quick synchronization among two processing steps. In general, according to the current technology state of art, latency issues may suggest considering the RANaaS centralization mostly for technologies at the control plane level. In fact, latency at data plane level can seriously undermine the transmission quality.

3.3.2 Functional constraints

This section briefly explore the most relevant implications of using virtualization and cloud computing for supporting iJOIN candidate technologies.

Virtualization can be defined as a technique aimed to simulate the existence of a piece of hardware actually "materialized" by a software layer running on top of the physical device. The idea is that the actual hardware is hidden to the applications and it is partially or temporary used for "impersonating" the role of a virtual piece of similar hardware.

For example, server virtualization creates the illusion of running several (virtual) servers on top of a single physical computer; this is obtained using a hypervisor (a.k.a.Virtual Machine Monitor or VMM) that is a piece of software tightly integrated with the operating system installed on the physical server.

The following figure (Figure 3-11) summarizes the concept.

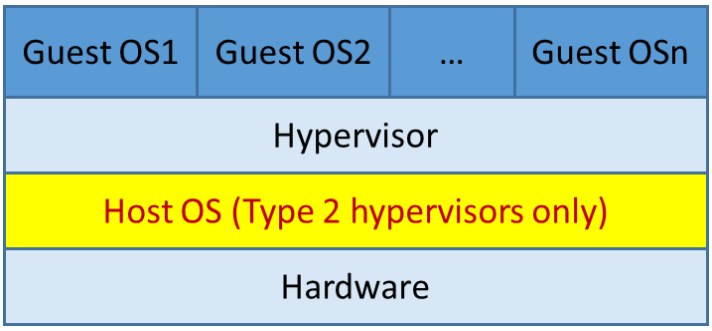


Figure 3-11: Server Virtualization

At the bottom of the stack, the physical hardware provides the actual computational resources (e.g. CPUs and RAM). In case of Type 2 hypervisors (e.g. Microsoft Hyper-V) , an operating system - called Host Operating System - is installed on the bare metal and it is integrated with the hypervisor, the component responsible of creating virtual servers. In case of Type 1 hypervisors, the hypervisor is directly running on the system hardware, without being integrated with a hosting OS: this is the case, for instance, of VMWare ESXi, XEN or KVM. In both cases, different operating systems called Guest Operating Systems run on top of the hypervisor. Each virtual server appears as an autonomous computer having its own (virtual) hardware; users access virtual servers via network connections.

Hypervisors are designed for having little impact on processing power and, most of the times, concurrent processes/threads[3] on the guest servers have performances similar to concurrent processes/threads in a more traditional time-sharing system without virtualization. In addition, in case a VM is assigned a certain number N of (virtual) CPUs, it can (virtually) execute up to N processes/threads in parallel. It is important to mention that when a VM is started, it is possible to define the number of virtual CPUs that the VM will use. This number defines a virtual parallelism that becomes real parallelism when the number of physical CPUs dedicated to the execution of the VM corresponds to the number of real CPUs dedicated to it. This aspect is regulated by the overbooking factor.

Overbooking can be defined as the ability of running, on a physical server, a number of virtual servers requiring more hardware resources than the ones available. For example, a physical server with 8 CPUs can run a certain number of virtual machines allocating a total of 10 virtual CPUs (i.e., vCPUs). This is possible because normally not all the virtual machines are running at the same time; therefore, when a VM is waiting for a "slow event" (i.e. an interrupt), the real CPUs are used for running concurrent VMs.

Similarly a physical server with 64 GB RAM can accomodate a number of virtual machines requiring 200 GB RAM; in such a case, memory swapping techniques are used for transfering main memory "chunks" on the mass storage. Overbooking can have strong impact on the system performance because it may happen that, under certain conditions, the actual workload of a physical server cannot be supported by the available physical resources and some VMs are randomly suspended independently on the priority of the applications they are running. In this case a simple software configuration can address the issue being sufficient to configure the virtualization software to prevent overbooking. This is an important aspect to consider when designing a real-time application and, specifically, parallel algorithms: actual parallelism is obtained by allocating enough virtual CPUs and disabling overbooking.

In addition, it is worth reminding that the amount of (virtual) resources allocated to the execution of a VM can only be defined at start-up and cannot be changed throughout the entire VM lifetime; consequently, when running a parallel algorithm on a single VM, it is fundamental to allocate a number of CPUs large enough for supporting the required level of parallelism. Then, if overbooking is disabled, the cloud computing platform (not the hypervisor) selects a physical server where this condition is satisfied and starts the VM on top of it.

Hypervisors introduce some overhead on interrupt management that results in higher latency with respect to situations where the processing is executed on 'bare metal' computing environment. This happens because when an interrupt occurs, the hypervisor must dispatch the related event to the 'right' VM, i.e. the VM that

[3] Processes and threads are two mechanisms that operating systems provide for implementing parallel programming.

was originally waiting for it. For example, when a network packet is received on a network interface, it must be dispatched to the VM that was waiting for it.

Figure 3-12 shows how interrupts are managed in a virtualized environment.

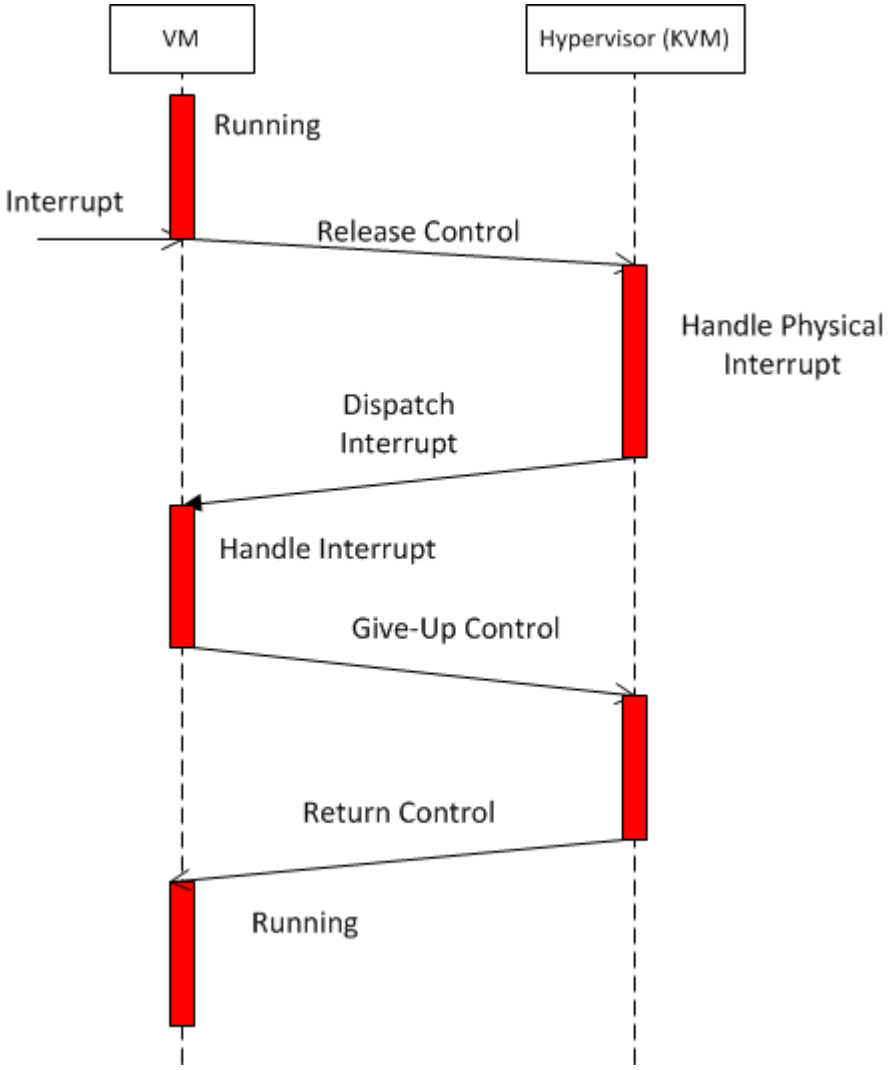


Figure 3-12: Interrupt Management in a virtualized environment

When an interrupt occurs, the currently running VM is halted and control is given to the hypervisor for identifying the VM originally waiting for the interrupt to occur (for the sake of simplicity, in this example there is only one VM running). Then the control is passed to the VM interrupt routine that, after managing the interrupt, returns control to the hypervisor. Eventually, the hypervisor decides what is the VM to return control to. The overhead introduced by the hypervisor can become substantial for I/O intensive applications, like iJOIN CTs, where hundreds of interrupts occur in a second.

Figures presented in [51] show that the typical latency to a message signal interrupt (MSI) on a virtualized environment with KVM hypervisor ranges from 300 to 700 µsec against a typical latency of 20 µsec on non-virtualized environments (Intel® Romley Server with 2 Intel® Xeon® CPU E5-2697 v2 processors,70GHz 12 cores x86_64 architecture):

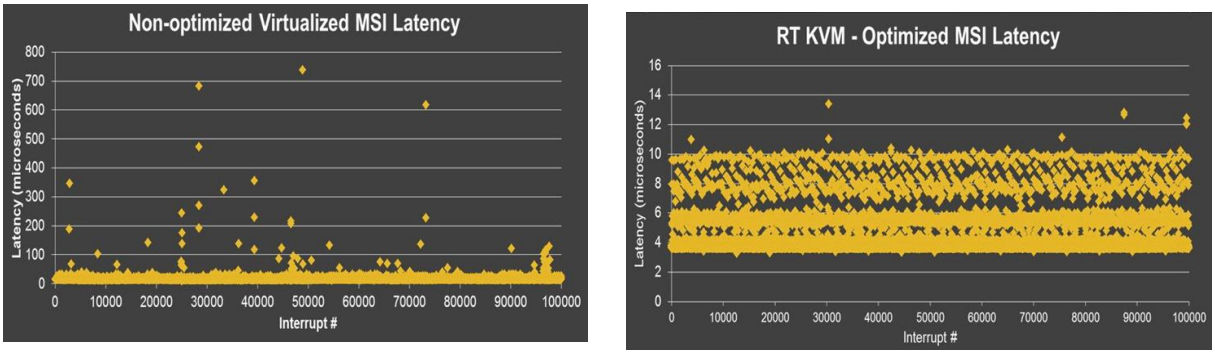


Figure 3-13: Non-optimized vs. optimizated virtualized MSI latency according to [67]

The situation described in Figure 3-13 (left) shows a significant degradation in performance, latency and determinism.

In order to address the problem, solutions have implemented or are going to be implemented in order to achieve *near-native* (i.e. similar to non-virtualized) performance. Broadly speaking, they consist in modifications of the virtualization software (e.g. KVM hypervisor) and/or of the underlying hardware platform aimed to select the 'best' hardware resources and exclusevely dedicate them for the processing of real-time workload hosted in a VM [51], [67].

Implementation of optimization mechanisms decribed in [67], shows significant improvements in the virtualized platform achieving results comparable to a non-virtualized (i.e. native) environment. Figure 3-13 (right) shows a maximum interrupt latency of less than 14 μs with average value around 8 μs, similar to the non-virtualized native interrupt latency of about 10 μs for the worst case and 3 μs average, as shown in Table 3-6.

Table 3-6 Interrupt latency (based on [21])

Test Conditions\	Interrupt Latency	
	Maximum (μs)	Average (μs)
No virtualization (native)	9.8	3
Optimized, virtualized	16.9	3.8
Non-Optimized, virtualized	760	25

One of the most important aspects and a promise of cloud computing is elasticity. Elasticity can be defined as the ability of a system to adapt to the workload changes by increasing/decreasing the amount of computing resources dedicated to an application. Given a certain application, when the workload increases, more computing resources are allocated; on the contrary, when the workload decreases, the number of computing resources is reduced accordingly. Elasticity aims at matching the amount of resources to the "real needs" in order to avoid over-provisioning or under-provisioning phenomena. In cloud computing, elasticity is implemented by provisioning or releasing virtual machines following a so called *scale-out* paradigm (or horizontal scalability) that spreads the executed workload over more computational nodes.

Generally speaking, applications can take advantage of elasticity provided that we take into account the following inherent aspects:

- **Resource Provisioning Time**: starting a new VM may take up to several minutes and the start-up time depends on several factors such as the VM type, the image size, the datacentre operating conditions, etc;
- **Monitoring**: the workload must be constantly monitored and when the working conditions change (i.e. workload increases or decreses) the number of dedicated VMs shall change accordingly;
- **Coordination**: the workload must be distributed among all the VMs participating to the application assigning each VM a part of the overall processing.

In this perspective, cloud applications are usually implemented as sets of VMs collaborating for supporting the workload; some VMs are dedicated to pure processing (i.e. workers), others to monitor and coordinate. The latter interoperate with the cloud computing infrastructure for starting/stopping VMs when the workload conditions change.

In iJOIN context, programs implementing the CT functionality can be integrated inside a proper virtual machine image, so that whenever we need to activate a new instance of the CT, it is only a matter of creating and activating a new virtual machine configured with the right software image. The virtual machine comes with the due amount of virtual memory and virtual storage. If the CT algorithm has a distributed shape, this can turn into a multithread application running inside the virtual machine, as well as spreading the algorithm across multiple virtual machines.

For each algorithm (or individual algorithm thread), a dedicated bootable software image is prepared, including the baseline operating system plus the specific CT software. The image is then duly configured to interact with the OpenStack platform. This turns into a set of operations, encompassing for instance all the network configuration (removal of hardcoded MAC addresses, enabling DHCP,…) and the installation of the software chunks needed to make OpenStack control the virtual machine. Then the software image is loaded into the OpenStack catalogue.

The realized images can then be grouped into *RAN services*, logical bundles including a set of images, network configuration and storage virtual volumes defining a self-consistent instance of a RAN functionality. These RAN services could be regarded like an extension of the VNFs (Virtual Network Functions) concept, well known in the intra-datacentre network realm [52]. Every time we need to activate a new instance of the RAN service, OpenStack executes an orchestrated activation sequence, where the software images are instantiated into actual virtual machines, the network resources are created and configured, etc. A very simple example of instantiation trigger could be the switch-on of an iJOIN small cell from a standby state. With the same mechanism, virtual resources can be released once for any reason we decide to stop a given instance of the RAN functionality (e.g., due to a traffic downturn we switch off an active iSC).

iJOIN does not envision a monolithic, full-fledge porting of the 3GPP LTE stack into the veNB; on the contrary, iJOIN targets a modular and even dynamic environment, where, according to the actual constraints and convenience, only the best suited part of the stack functions is executed into the veNB. To duly follow this execution model, at least at initial stage every CT must be implemented into its own virtual machine, to keep the ability of autonomous execution and flexible centralization of candidate technologies (Figure 3-14). In turn, each virtual machine can be instantiated several times, so that each CT can be scaled out based on the dynamic conditions of capacity demand.

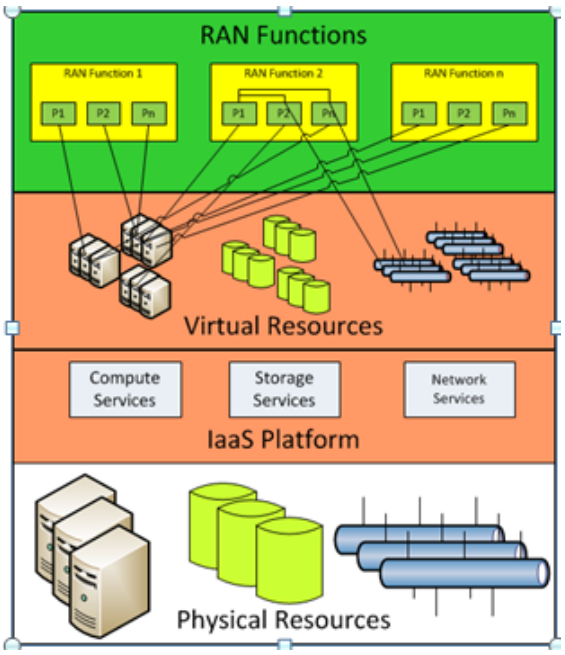


Figure 3-14: Mapping of CTs to virtual machines

The communication among virtual machines occurs through *virtual networks*. Network resource virtualization is fully alike the similar virtualization concept for server or storage resources: underlying

network links are assigned to virtual machines as if they were exclusive physical links, and the hypervisor takes care of handling the virtual/physical translation. On top of these virtual links, typical network communication mechanisms are used (e.g. TCP/IP sockets). It is important to underline that as "network" we mean here the *internal* interconnections of the datacentre (or the datacentres) hosting the RANaaS cloud platform, not the mobile network or the backhaul.

4 iJOIN MAC/RRM Candidate Technologies

4.1 CT 3.1: Backhaul Link Scheduling and QoS-aware Flow Forwarding

4.1.1 Technical description

Scenario

This CT considers a dense small cell network deployment where RANaaS can be seen as a coordinator and traffic aggregator for a cluster of iSCs. The small cell backhaul (iSC-iSC, iSC-RANaaS) is considered to operate in millimetre wave radio (60GHz) to provide high bandwidth and low latency data transfer which can be comparable to wired backhaul.

In the following figure the system model is shown which consists of RANaaS for the cluster of small cells. Due to the high path losses in 60GHz, a number of hops might be required from RANaaS to reach all the destination iSCs.

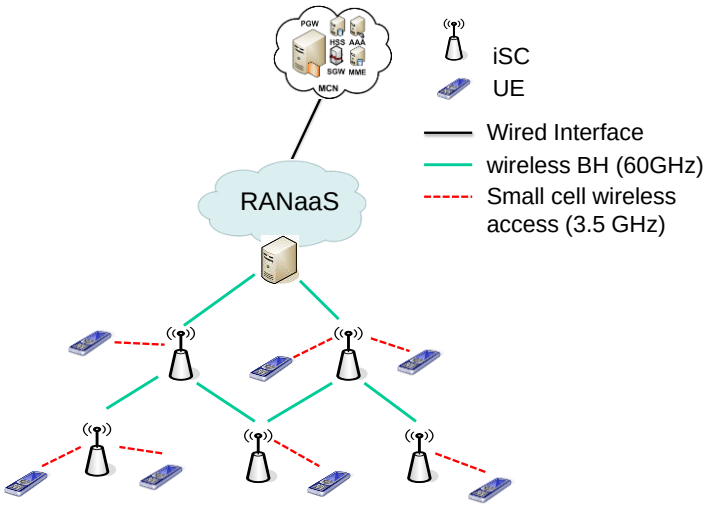


Figure 4-1: Backhaul link scheduling and QoS-aware flow forwarding

In the multi-hop wireless BH consisting of small cells, routing information is required at the RANaaS entity from network layer so as to select paths (in larger time-scale) based on traffic load, energy efficiency and other criteria. Taking into account this information as an input from the network layer (with the aid of the already defined iNC from WP4) and the channel conditions between iSCs, our work has the following objectives:

Firstly, to dynamically identify BH links and paths to be scheduled per a given time window, taking into account the target global objective for the network (in terms of maximizing backhaul capacity or aggregate utility). Here, BH link scheduling is performed in a centralized way (RANaaS).

Secondly, assuming we have heterogeneous traffic (real time, non-real time), to identify how the incoming flows are stored in the queues and forwarded to the next hops (or destinations), taking into account the link selections in the previous step and the fulfilment of the QoS requirements (delay, outage and data rate) per flow.

System Model

As can be seen in Figure 4-1, the system consists of a RANaaS entity, which serves as a controller and traffic aggregator for a dense small cell network. This small cell network encloses $l=1,2,..,L$ iSCs, equipped with antennas for the access (iSC-to-user), as well as directional antennas for the small cell backhaul (which operate in higher frequencies). The latter is essential to exchange signalling and data with RANaaS (or other iSCs), using wireless backhaul (millimetre wave).

Let $G(V,E)$ the graph consisting of a set of V nodes (iSCs) and a set of E edges. An edge $e \in E$ is a connection between any two nodes $v1, v2 \in V$. The edge $e \in E$ indicates that data can be exchanged between $v1$ and $v2$. We assume we have m links in the network, where the links are considered un-

directional. Here we introduce *t(e)* and *r(e)* as the transmission and receiver ends of edge $e \in E$ respectively. Using this, we define the sets $N_{input}(v)$ and $N_{output}(v)$ as the sets of links which terminate and originate to/from node *v* accordingly.

Here we also introduce a set of *M* demands where the m^{th} demand originates from the source node s_m and terminates to the destination node d_m with a required rate r_m and maximum delay τ_m. Each link *e* has a desired data rate which corresponds to the summation of all the demands traversing it. We define the load of link *e* due to demand *m* as *l(e,m)*. Hence, the total load of this link is defined as:

$$l(e) = \sum_{m \in M} l(e,m) \tag{4.1}$$

Another important parameter is the link capacity, defined as $c(e), \forall e \in E$. Here, high directional antennas can be used to compensate for the high path attenuation. In this case, interference by other links is assumed to be negligible due to the high directivity of the antennas and the half duplex constraint (nodes only transmit or receive). Hence, the channel is only affected by the path attenuation. It is high probable that we may have LoS between iSCs; however there is the chance that LoS is not available for an unplanned dense iSC deployment.

As can be seen below, *c(e)* can be computed using the LoS probability P_l , where *c(e, LoS)* is the link capacity having LoS (free space loss) and *c(e, nLoS)* is the case when we have non LoS. Here we introduce Bernoulli random variable *I(x)* based on the probability of LoS ($x=P_l$).

$$\mathrm{c}(e) = \mathrm{I}(\mathrm{P}_l)\,\mathrm{c}(\mathrm{e},\mathrm{LoS}) + (1 - \mathrm{I}(\mathrm{P}_l))\,\mathrm{c}(\mathrm{e},\mathrm{nLoS}) \tag{4.2}$$

Using the definitions of the link capacity and the link load, we can introduce a new parameter which captures the number of time-slots required for a link to satisfy this demand.

$$f_e = \left\lceil \frac{\sum_{m \in M} l(e,m)}{c(e)} \right\rceil \tag{4.3}$$

Here to mention that in case this ratio is fractional, we round the value to the next integer, since we aim to find the number of time instances required.

Another key parameter for the scheduling part is the set of all the bi-partite sub-graphs of the graph *G(V,E)*, denoted as $\mathcal{S}$. Each of these S_i sub-graphs represents a combination of link activations (one set of the bi-partite graph is the transmitter nodes and the other set is the receiver nodes). Each of these sub-graphs is associated with a weight factor w_{Si} which represents the fraction of time that this combination of activations is active. Below we illustrate simple example of two different combinations for 6 random nodes. Note that the total number of these combinations is $2^{|V|} - 2$. We also define a binary indicator variable $1_{e,S}$ which is 1 if the edge e is included in sub-graph S_i (otherwise 0).

The joint path selection and scheduling problem can be written as the following optimization problem. The maximization of total BH throughput is equivalent to the minimization of the total number of timeslots, which define the ratio of the demand over the backhaul link capacity towards an iSC. In other words, the objective is to find which routes the traffic should follow and which links to be active so as to maximize performance.

$$\min \sum_{e \in E} f_e x_e \tag{4.4}$$

Subject to:

$$\sum_{e=\{0,j\} \in E} x_e = k \tag{4.5}$$

$$\sum_{e=\{\mathrm{i},j\} \in E} x_e \le 1, \forall i \in V \tag{4.6}$$

$$\sum_{e=\{j,i\}\in E} x_e = 1, \forall i \in V \tag{4.7}$$

$$\sum_{e=\{i,j\}\in S_k} f_e x_e \le D_{\max}, \forall k \tag{4.8}$$

$$f_e \le T \sum_{i=1}^{|B|} w_{B_i} 1_{e,B_i}, \forall e \in E \tag{4.9}$$

$$\sum_{i=1}^{|B|} w_{B_i} = 1, w_{B_i} \ge 0, \forall B_i \in B \subseteq S \tag{4.10}$$

The first 4 constraints (4.5) – (4.8) are the routing constraints, whereas constraints (4.9), (4.10) are the scheduling constraints. In (4.5), the number of links between RANaaS (denoted as node 0) and all the iSCs depend on the number of routes and is equal to the variable *k*. The highest we set this value, the lower hops are expected in total. In (4.6) and (4.7), the number of incoming edges and outgoing edges to/from each iSC is set exactly of less than one. By this, all the iSC must be able to receive traffic and at the same time it is optional to have outgoing traffic to other links. Constraint (4.8) is the maximum delay constraint which has to be taken into account when creating a route. This constraint might be variable depending on the traffic (i.e. low threshold for real time, high for non-real time traffic). Moreover, (4.9) shows that the cost of the link shall not exceed the pre-defined time window (*T*); and finally (4.10) shows that the summation of the weights (which show the fraction of time each sub-graph is active) is set to one.

Approach

The problem in (4.4) is a NP-hard combinatorial optimization problem. Therefore, our proposed framework decouples the initial problem in two sub-problems. At the first stage, we target to solve the path selection problem (constraints (4.5) –(4.8)) which has the form of an Integer Programming problem. By this, we identify which links to activate and how many slots to dedicate to this links, such that the BH throughput is optimized. The solution of this sub-problem is found using the Branch-and-Cut exact approach. The next stage is the selection of the packet forwarding from the queues in a way that the delay is minimized, taking into account the half duplex constraint, the multi-hop requirements and the queue buffers. This problem is solved using a proposed back-pressure scheduling algorithm.

1) Path Selection Algorithm

The objective of this problem is to deliver to a set of iSCs with known traffic demands on minimum cost routes originating from RANaaS (given a pre-defined number of routes *k*). As discussed above, this is an Integer programming problem that can be solved by Branch-and-Cut algorithm.

The algorithm follows a branch-and-bound scheme, where lower bounds are computed by solving a linear program (LP) relaxation of the problem. This relaxation is iteratively tightened by adding valid inequalities to the formulation according to the cutting plane approach. The exact method is known as a branch-and-cut algorithm and is thoroughly described in [54] for the case of the IP problem. Following, we briefly describe the algorithmic steps we used in our study:

a) Initialization: At this stage we transform the initial graph to an edge graph so as to be able to solve the IP problem. The resulting edge graph which includes the iSC-iSC, iSC-RANaaS potential links, defines the number of variables in the IP problem.

b) Lower Bound: Having formed the edge graph, the next step is to find the lower bound using an LP relaxation. In our work the initial near-optimal solution for the root node is derived using Langragian relaxation.

c) Upper Bound: After finding the lower bound, which is the optimal solution for the relaxed problem, we now aim to find the upper bound to the original problem, which is a set of feasible solutions using local search algorithms and improvement procedures, in similar way as in [54].

d) Branching: Here, we create a new node in the search tree following the logic of branch and bound. We consider the branching on variables, the standard approach for branch-and cut. It consists of selecting a

fractional edge-decision variable and generating two descendant nodes by fixing its value to either 0 or 1. In our implementation, we use the most fractional branching where we choose variable with fractional value closest to 0.5 (ties are broken by choosing the edge having maximum cost)

2) Scheduling Algorithm

After obtaining the routes and the number of timeslots that each link is going to be used for all destinations, the next phase is to find how to forward the packets from RANaaS to all the iSCs, having a variable number of hops per route with the minimum delay.

The packets are stored in separate queues per destination at the traffic aggregator. The target is to empty all the queues by the end of a given time window. Here, the constraints are that at each time-slot, one node can only transmit or receive packets to one destination (half duplex constraint). Furthermore, the traffic that is forwarded through more than one hop must be stored in separate queues in the intermediate nodes. In each queue FIFO policy is applied and also there is a threshold for the highest number of packets that can be stored in each queue.

In Figure 4-2, an example can be shown with 1 starting node (RANaaS) and 4 iSCs. The first two iSCs will require data for their users, as well as for transferring data to iSCs 3 and 4 respectively. The routes to be followed and the timeslots needed are known from the previous stage. Hence, the problem is to find how to forward the packets so as to empty all queues in the minimum time.

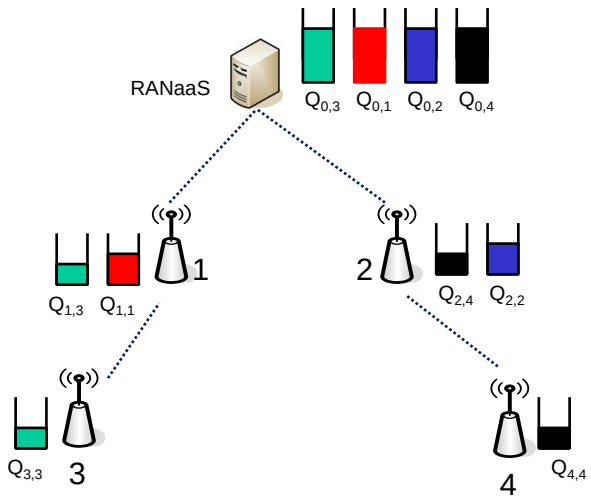


Figure 4-2 Back-pressure scheduling example for 4 iSCs

For the solution of this problem we propose a throughput optimal algorithm which follows the back-pressure concept [55]. Assuming slotted time, the basic idea of backpressure scheduling is to select a set of non-interfering links for transmission at each slot. Non-interfering links refer to links that do not have the same transmitting and/or receiving end, such that the half duplex constraint is maintained. Here, the objective is to serve the flows *f* with the maximum differential backlog. The differential backlog for each node *i,j* is defined as $\Delta Q_{i,j}^{f} = Q_i^f - Q_j^f$. The steps of this algorithm are the following:

- *Step 1*: Compute the weight of each link *(i,j)* as $w_{i,j,t} = \max_f \left(Q_{i,t}^f - Q_{j,t}^f\right)$
- *Step 2*: Select links to maximize: $x^*(t) = \arg\max_x \sum_{(i,j)} w_{i,j,t} x_{i,j,t}$, where $\sum_t x_{i,j,t} = x_e : \mathrm{e} = \{\mathrm{i}, \mathrm{j}\} \in \mathrm{E}$
- *Step 3*: Transmit the chosen flows on the selected links

For the example shown in Figure 4-2, we observe that there are k=2 routes (0-1-3 and 0-2-4). The number of flows are *k(k-1)/2* =6 flows. So, we measure each timeslot the differential backlogs and forward the packets for the links that maximize it. The algorithm stops where no packets are left at the queues. Here to mention that during the scheduling phase, no more packets are assumed to arrive at the traffic aggregator.

4.1.2 Implementation of CT in the iJOIN architecture

The message sequence chart for the proposed scheme can be shown in Figure 4-3. RANaaS first receives BH routing tables/info for the available routes towards all iSCs which are scheduled in larger time scale. RANaaS also receives channel state information for the BH / access channels and the queue buffer status of each iSC. Taking into account these inputs, the RANaaS entity schedules the link activations and forwards the data through one or more hops. For every exchanged parameter, the corresponding identifier (I3.x, O3.x) is given as defined in D3.1 [5] and summarized in Appendix I.

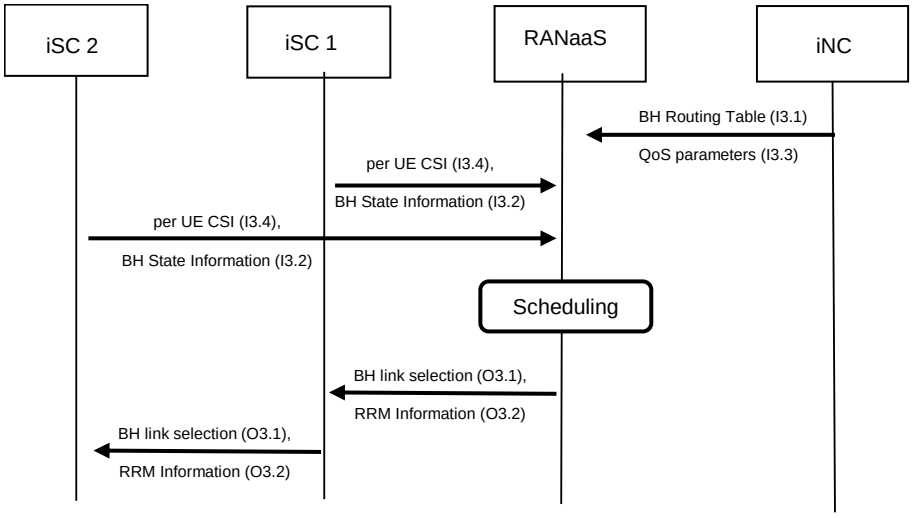


Figure 4-3 Message Sequence Chart for CT3.1

4.1.3 Evaluation of the CT

The evaluation of this CT was performed using system level simulations for the Wide Area Scenario (CS3). Here, 19 iSCs are controlled by a RANaaS entity, as illustrated in Figure 4-4.

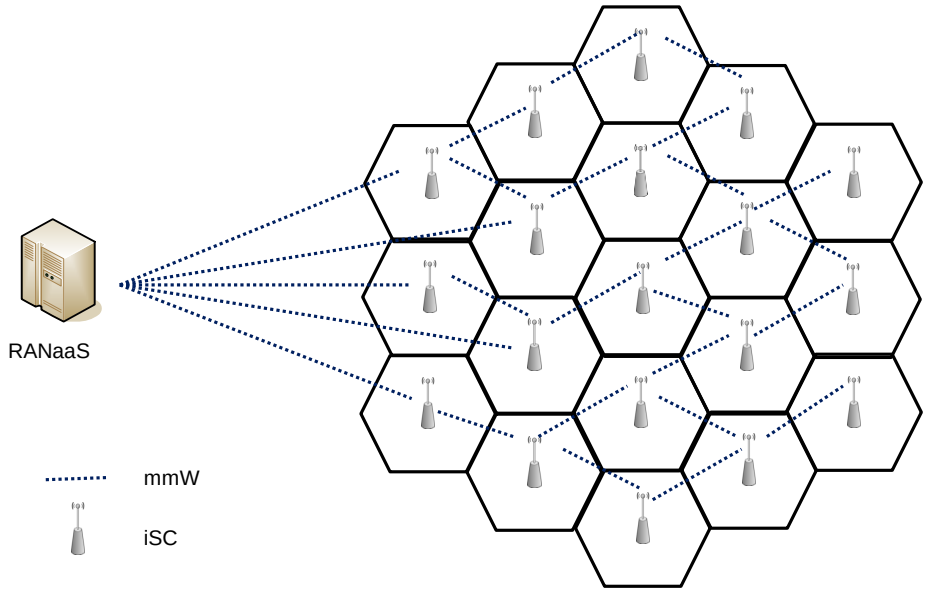


Figure 4-4 Small Cell Deployment for CT3.1

Two channel models are used for the 60GHz BH: LoS channel (free space loss) and NLoS channel using an empirical path loss model from [56]. For each link, a LoS probability function was used from literature [57] to identify whether a link in LoS or nLOS.

Compliance with iJOIN objectives

This CT proposes the efficient BH link scheduling (activation / de-activation) in a millimetre-wave small cell BH environment so as to ensure high capacity and low latency small cell backhaul taking into account the traffic demand for the access per iSC and the users' QoS requirements for different types of traffic.

Description of the baseline used for the evaluation

The proposed candidate technology is compared against different multi-hop scenarios to better capture the trade-off between delay and BH link throughput. The evaluation will also enclose other BH technologies to better capture the realistic gain when using mmW BH in a multi-small cell environment.

Discussion of results of the CT

The implementation of this scheme had two parts. The first part is the extraction of results for the path selection problem. Here, we were adjusting the number of routes (*k*), so as to find the optimal path selection using different number of paths. In Figure 4-5, two extreme cases are shown. The first case is the single-hop case (*k*=19, Figure 4-5-left), and the second case of having 1 route which includes 19 hops (k=1, Figure 4-5-right).

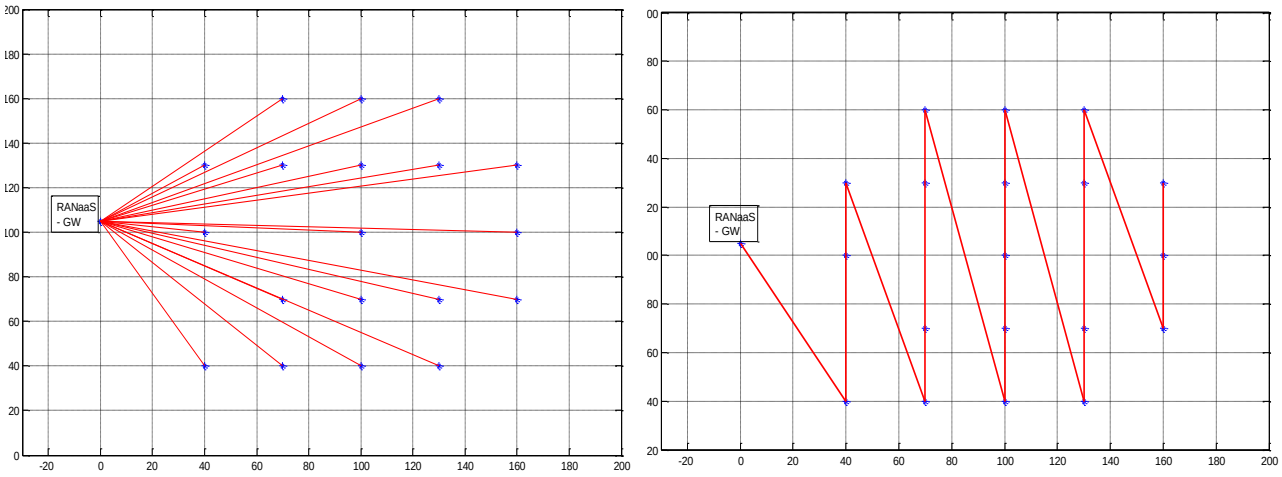


Figure 4-5 Illustration of BH topology for k=19 and k=1

Then, the path selection algorithm was tested for all the possible number of paths. Below, in Figure 4-6 we observe that the average BH link spectral efficiency drops when we increase *k*. This is due to the fact that the higher the number of routes, the lower the number of hops we have. So, we might have long- distanced links with NLOS which can affect the performance. On the other hand, low *k* means more hops with less routes; hence the possibility of more short-distanced LoS hops can increase performance.

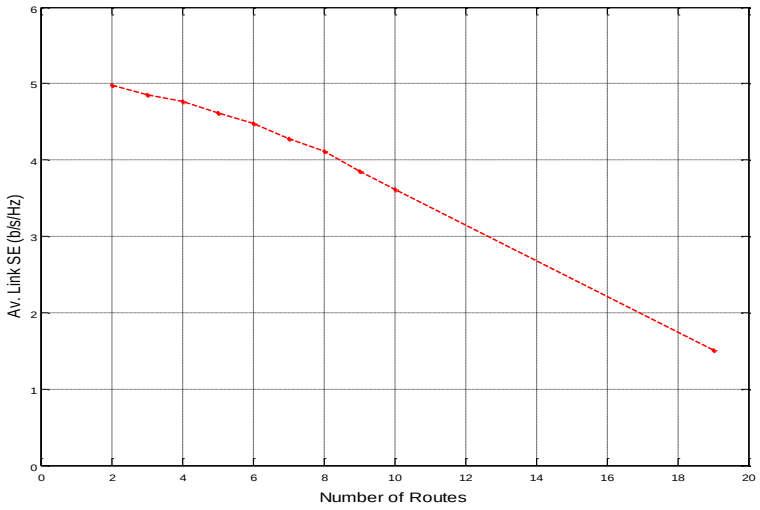


Figure 4-6 Average BH link spectral efficiency vs. number of routes

However, this is a trade-off since the high number of hops might increase the delay. In the second part of our implementation the backpressure scheduling was used. As can be seen in Figure 4-7, we extract the maximum delay (the number of timeslots until the last iSC receives the last packet) and the average delay (the average number of timeslots till each iSC is served for its access). We observe that the higher the

number of routes the lower the delay for both maximum and average curves. This shows that we may achieve higher throughput with more routes; however this comes at the cost of higher delays. Another important comparison in this figure is the evaluation of different delay threshold. This delay threshold *Dmax* was discussed in (4.8). We set two different values for this threshold (low for real-time and high for non-real time) and we observed that the delay having lower threshold gets lower as the number of routes increases.

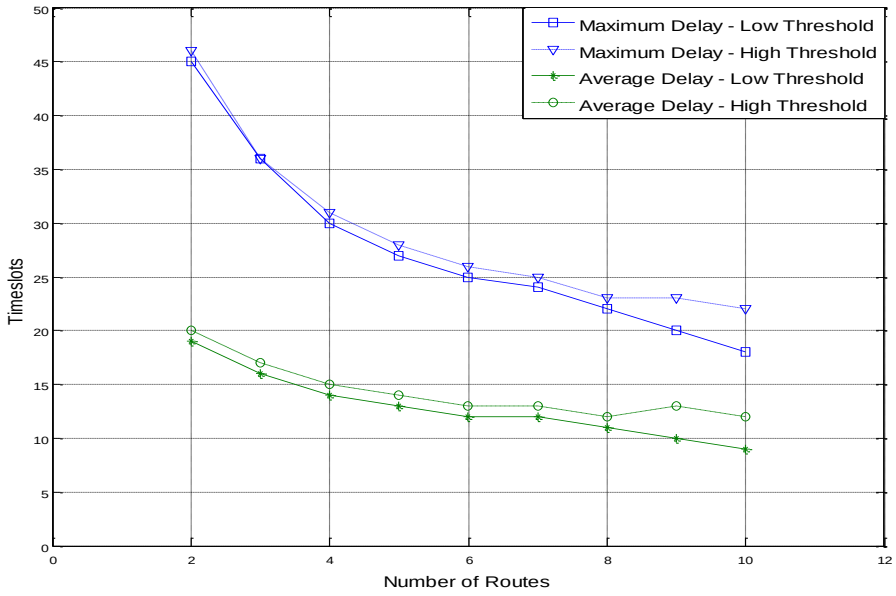


Figure 4-7 Maximum and average delay for different delay bounds vs. number of routes

Finally in Figure 4-8, the trade-off between maximum delay and average link throughput is shown. Here, the higher the number of routes, the lower the spectral efficiency and the delay.

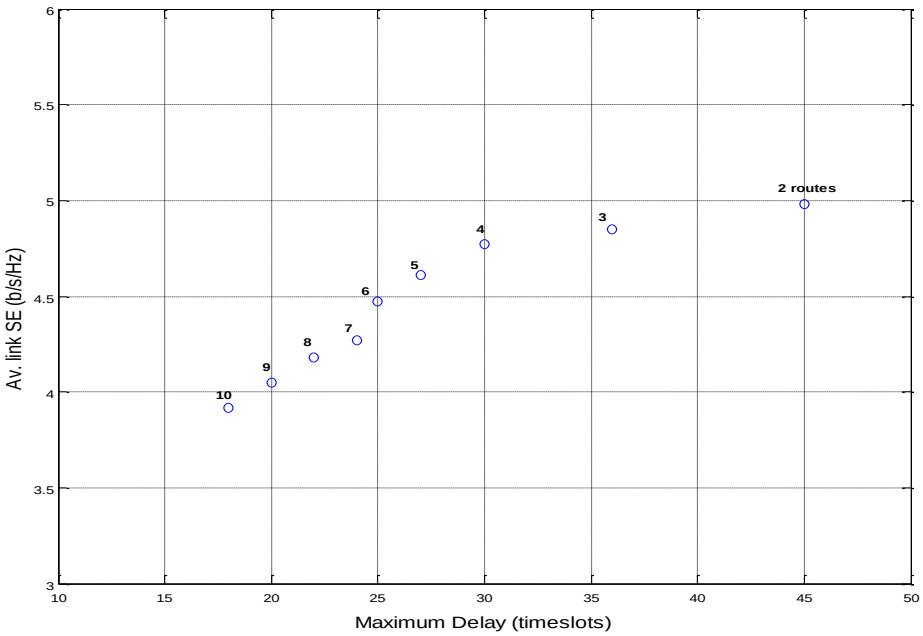


Figure 4-8 BH link Spectral Efficiency vs. Maximum Delay

4.2 CT 3.2: Partly decentralized mechanisms for joint RAN and backhaul optimization in dense small cell deployments

4.2.1 Technical description

Scenario

This CT investigates a scenario where iSCs are densely deployed to satisfy the demand of high data rate services of future wireless networks. In this scenario, we will introduce backhaul aware cell selection mechanisms to enable network-wide load balancing and improve the overall network capacity.

We aim to investigate relationships between RAN capacity, cell load, resource scheduling (both at the backhaul and at the radio access), and backhaul capacity. Moreover, we aim to propose innovative cell selection mechanisms, where the above parameters are jointly considered.

System Model

We consider a mobile wireless cellular network in which user terminals and eNBs implement an OFDMA air interface based on 3GPP/LTE downlink (DL) specifications [2]. Coherently with the study on small cell enhancement, which is currently under investigation in 3GPP [3], our research focuses on HetNets where small cells are densely deployed and operate in a dedicated carrier with respect to the macro cell (see Figure 4-9). We also consider the presence of a controller, named as RANaaS [4], which orchestrates a cluster of small cells and connects this cluster with the core network. The RANaaS is connected to neighbouring Macro eNBs (MeNBs) through the X2 interface while the J1 interface is used to enable coordination amongst iSCs and the RANaaS.

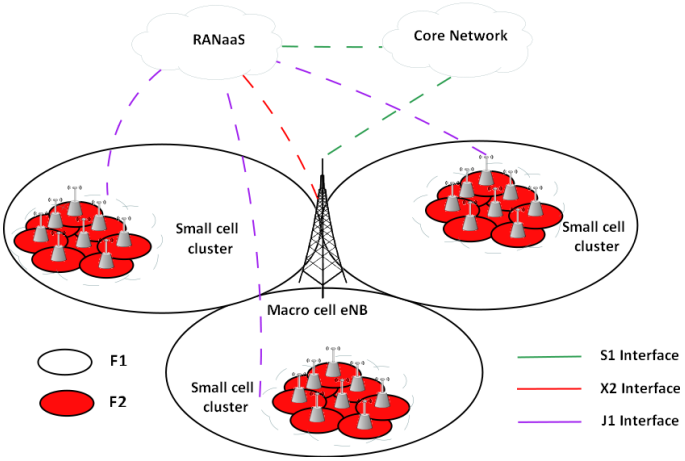


Figure 4-9: The heterogeneous network deployment under investigation.F1 and F2 are the carrier frequencies for the macro layer and the small cell layer, respectively.

In D3.1[5], we have presented our model that characterizes the relationships amongst cell load, backhaul capacity, and radio access network capacity. The objective pursued in this CT is finding the association amongst UEs and eNBs, α*, that maximizes the overall network capacity. The optimization problem can be simply expressed as

$$Find \quad \alpha^* = \arg\max_{\alpha} C(\alpha)$$

At a first glance, this combinatorial optimization problem may seem similar to a multiple knapsack problem [6], in which N items (the UEs) have to be associated to M knapsacks (the eNBs), each one of which has a limited weight (the capacity of the corresponding backhaul link), such that to maximize the profit (the overall capacity of the wireless network). In reality, our optimization problem is even more general than the multiple knapsack problem, since the UEs do not have a priori weight and profit, but these values are dependent on the association itself and on the resource allocation. Indeed, for each association α, each user *u* contributes to the weight of *α(u)* and to the value of the total profit *C(α)*, according to the quality of the link *(u, α(u))* and the resource allocation at *α(u)*. Since the knapsack problem is NP-complete, we also expect our optimization problem to be so, although a formal proof of such a result is out of the scope of this study.

Brute force algorithms, which evaluate all possible solutions and select the best one, might be used to solve simple combinatorial problems. According to our model, the size of the set *V*, which represents all the feasible solutions, can be computed as

$$|V| = \prod_{u}^{U} S(u), \tag{4.11}$$

where *S(u)* is the set that includes the eNBs in the active set of the user *u*.

Therefore, even in moderate dense deployment scenarios, computational/memory costs may prevent to find an optimal solution by using Brute Force (BF). Henceforth, in the following, we propose and investigate two iterative algorithms characterized by a limited complexity and designed to improve the overall network capacity by optimizing the cell selection process.

Centralized Approach

The proposed algorithm starts from a given simple solution of the cell selection problem, and evolves towards a more beneficial association. At each iteration, *Evolve* calculates and evaluates each possible change in the current association, and then selects the strategy which increases the most the overall network capacity. The algorithm stops after a limited number of iterations, when the achievable gain becomes less than a small non-negative value ϵ.

Let denote

- *G* the bipartite graph with vertices *U* and *S*, in which there is an edge between a user *u* and an eNB *s*, only if *u* is in the coverage area of *s* (i.e., SINR(*u, s*) $\geq \gamma_{th}$);
- *S(u)* = {$s \in S \mid (u, s) \in G$}, the eNBs in the active set of the user *u*;
- *U(s)* = {$u \in U \mid (u, s) \in G$}, the UEs located in the coverage area of *s*.

0. Initialization Step

- Let α be the state-of-the-art user assignment that associates to each user *u* the eNB *s* maximizing SINR(*u, s*), that is: $\alpha(u) = \arg\max_{s \in S(u)} SINR(u,s)$.
- For all $s \in S$, compute the associated capacity $C_\alpha(s)$ according to the used scheduler [7].
- For all $(u, s) \in G$, compute $X_\alpha(u, s)$, which measures the new capacity at the eNB *s* whether we change the association *α* by associating (respectively, de-associating) the user *u* to (respectively, from) *s*

$$X_\alpha(u,s) = \begin{cases} 0, \ if \ \alpha(u) = s \ and \ d_\alpha(s) = 1 \\ D_\alpha(s)^{\neg u}, \ if \ \alpha(u) = s \ and \ d_\alpha(s) > 1 \ and \ D_\alpha(s)^{\neg u} \leq C^{BH}(s) \\ D_\alpha(s)^{\oplus u}, \ if \ \alpha(u) \neq s \ and \ D_\alpha(s)^{\oplus u} \leq C^{BH}(s) \\ C^{BH}(s), \ otherwise \end{cases} \tag{4.12}$$

where $d_\alpha(s) = |U_\alpha(s)|$ is the number of users associated with *s*. The values of $D_\alpha(s)^{\neg u}$ and $D_\alpha(s)^{\oplus u}$ with respect to the different resource allocation policies are shown in

Table 4-1.

- For all $(u, s) \in G$, compute the gain $\Delta_\alpha(u, s)$ due to the possible reassignments of the user *u* from the eNB α(*u*) to the eNB *s*:

$$\Delta_\alpha(u,s) = \begin{cases} X_\alpha(u,\alpha(u)) + X_\alpha(u,s) - C_\alpha(\alpha(u)) - C_\alpha(s), & if \ \alpha(u) \neq s \\ 0, & otherwise \end{cases} \tag{4.13}$$

1. One-User Reassignment Step

- *Find* $(u_0, s_0) = \arg\max_{(u,s) \in G} \Delta_\alpha(u,s)$.

  Note that $\Delta_\alpha(u_0, s_0) \geq 0$, since $\Delta_\alpha(u, \alpha(u)) = 0 \ \ \forall u \in U$.

- *If* $\Delta_\alpha(u_0, s_0) \leq \varepsilon$ *exit* (the algorithm outputs the current assignment α).

Table 4-1: $D_\alpha(s)^{\neg u}$ and $D_\alpha(s)^{\oplus u}$ with respect to different resource allocation policies. *η(u,s)* represents the spectral efficiency between u and s.

	Round Robin	**Data Rate Fairness**	**Max C/I**
$D_\alpha(s)^{\neg u}$	$\frac{B}{d_\alpha(s)-1} \cdot \sum_{u' \in U_\alpha(s)-u} \eta(u',s)$	$\frac{B \cdot (d_\alpha(s)-1)}{\sum_{u' \in U_\alpha(s)-u} \frac{1}{\eta(u',s)}}$	$B \cdot \frac{\sum_{u' \in U_\alpha(s)-u} \eta(u',s)^2}{\sum_{u' \in U_\alpha(s)-u} \eta(u',s)}$
$D_\alpha(s)^{\oplus u}$	$\frac{B}{d_\alpha(s)+1} \cdot \sum_{u' \in U_\alpha(s) \cup u} \eta(u',s)$	$\frac{B \cdot (d_\alpha(s)+1)}{\sum_{u' \in U_\alpha(s) \cup u} \frac{1}{\eta(u',s)}}$	$B \cdot \frac{\sum_{u' \in U_\alpha(s) \cup u} \eta(u',s)^2}{\sum_{u' \in U_\alpha(s) \cup u} \eta(u',s)}$

- Define $s_* = \alpha(u_0)$ (hence $s_* \neq s_0$)
- Define a new user assignment α_0 by:

$$\alpha_0(u) = \begin{cases} \alpha(u), & \text{if } u \neq u_0 \\ s_0, & \text{if } u = u_0 \end{cases} \tag{4.14}$$

2. *Metric Update Step*

- For all $s \in S$,

$$C_{\alpha_0}(s) = \begin{cases} C_\alpha(s), & \text{if } s \neq s_* \text{ and } s \neq s_0 \\ X_\alpha(u_0, s), & \text{if } s = s_* \text{ or } s = s_0 \end{cases} \tag{4.15}$$

- Set $X_{\alpha_0} = X_\alpha(u,s)$, for all $s \in S \setminus \{s_*, s_0\}$ and $u \in U(s)$; Compute $X_{\alpha_0}(u,s)$ for $s \in \{s_*, s_0\}$ and $u \in U(s)$.
- Set $\Delta_{\alpha_0}(u,s) = \Delta_\alpha(u,s)$ for all $s \in S \setminus \{s_*, s_0\}$ and $u \in U(s)$; Compute $\Delta_{\alpha_0}(u,s)$ for $s \in \{s_*, s_0\}$ and $u \in U(s)$.
- Set $\alpha = \alpha_0$, then go to Step (1).

The proposed framework for backhaul-aware cell selection is illustrated in Figure 4-10.

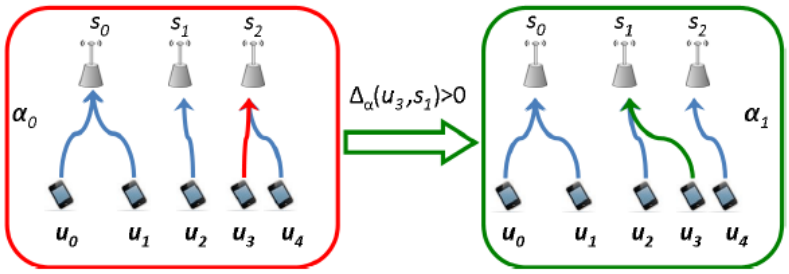


Figure 4-10: Evolve paradigm for managing user-eNB association.

Proposition: *In Evolve, the value of C(α) is improved at each new iteration. Hence, the algorithm converges when it is no possible to further improve the value of C(α) by a new reassignment of one single user.*

Proof: Let α be the current user assignment, possibly after some number of iterations. Let α_0 be the new reassignment, computed at Step (1). Then

$$C(\alpha_0) - C(\alpha) = \sum_{s \in S} C_{\alpha_0}(s) - \sum_{s \in S} C_\alpha(s) = C_{\alpha_0}(s_*) - C_{\alpha_0}(s_0) - C_\alpha(s_*) - C_\alpha(s_0) =$$

$$X_\alpha(u_0, s_*) - X_\alpha(u_0, s_0) - C_\alpha(s_*) - C_\alpha(s_0) = \Delta_\alpha(u, s) > \varepsilon$$

In particular, Evolve can guarantee at least the same performance as the SINR-based approach.

Distributed Approach

The second algorithm proposed is based on game theory. Game theory offers an interesting perspective to deal with distributed solutions, which achieve near optimum performance. Since these distributed solutions are less intensive in terms of computational load than the brute force calculation of the optimum cell selection scheme, we can use them as well to obtain an approximation to the optimal allocation in a fast way. Therefore, we model the cell selection process as a formal game and perform the algorithmic design by correctly defining the set of players and the utility functions.

A game is defined by the tuple $\Gamma = \{P, \{S_i\}_{i \in P}, \{u_i\}_{i \in P}\}$, where P is the finite set of players, S_i is the set of strategies of player i and u_i: $S \rightarrow$ R is the utility function of that player, with $S = \times_{i \in P} S_i$ the strategy space of the game.

The utility function u_i is a function of s_i, the strategy selected by player i, and of s_{-i}, the current strategy profile of the rest of the players of the game. Players will selfishly choose the actions that improve their utility functions considering the current strategies of the other players.

In our case, the players of the game represent the users that connect to the network ($P = U$). It is worth noting that it is not the physical users the ones that select their strategy in the game (i.e. physical users are not in charge of the cell selection process), but the network itself using a "virtual" representation of them as players of the game used to represent the cell selection process. Additionally, the set of strategies S_i of a user with n small cells is the set of small cells the user can connect to.

As for the utility function, one general key issue when designing a game is the choice of u_i so that the individual actions of the players provide a good overall performance. In addition, in our specific scenario it is interesting the existence of an equilibrium point to ensure the convergence of the proposed algorithm. In this context, it is useful the concept of Nash Equilibrium (NE), defined as a situation where no player has anything to gain by unilaterally deviating. Thus, a NE of a game Γ is a profile $s^* \in S$ of actions such that for every player $i \in P$, we have that $u_i(s_i^*, s_{-i}^*) \geq (s_i, s_{-i}^*)$ for all $s_i \in S$, where $s_{-i}^* \in S$ denotes the strategies of all players other than player i in the profile s^*.

From a radio resource allocation perspective, the convergence to a NE of the game makes it possible to reach a stable solution. In addition, the network can react to variations in the environment as any deviation from this equilibrium forces the participants to play again to lead a new NE. In our case, we consider three possible definitions for the utility function, given to the following three games:

Rate Local Game: This solution models each user as a selfish player, which wants to maximize its own capacity, but adding a certain grade of cooperation to approach the global maximization objective. The utility function of player *i* is directly related to the capacity of that user *i* achieves when it connects to the small cell corresponding to strategy s_i:

$$u_i\left(s_i, s_{-i}\right) = \begin{cases} C(u) & \text{if } C(u) > C_{\min} \text{ and } p_i > 0 \\ 0 & \text{if } p_i = 0 \\ -1 & \text{otherwise} \end{cases} \tag{4.16}$$

Therefore, the utility is the capacity of a specific user if and only if this capacity is higher than a predefined threshold $C_{\min}$. The value -1 in the utility function also tries to introduce a degree of cooperation to compensate the inherent selfishness of this game; if the user cannot be connected with the minimum predefined quality $C_{\min}$, it is better to stop its transmission to reduce the interference on the remaining users. To ensure the existence of at least one Nash Equilibrium (NE) and the convergence of the game to one of them, we set a threshold for the maximum number of non-consecutive times that a player can choose a specific strategy. With this simple rule, the game has a NE at least; if all the players remove the strategies that exceed their corresponding thresholds without achieving a NE, ultimately the strategy space of the game will be formed by only one possible strategy for each player, which must be a NE of the game.

One key decision is how the capacity of the backhaul is shared among the different users, which are connected to a specific small cell. For this game, we are going to consider that the capacity allocated to each player is proportional to the capacity the player would experience if the backhaul restriction of the small cell is not considered. That is:

$$C(u) = \frac{C_{nb}(u)}{\sum_{u' \in S_j} C_{nb}(u')} C_b(S_j) \tag{4.17}$$

where $C_b(S_j)$ is the backhaul capacity of small cell S_j, $C_{nb}(u)$ is the capacity that user *u* would get without backhaul restrictions (ideally, the Shannon capacity corresponding to its SINR) and *u'* are the different users connected to small cell S_j.

Rate Potential Game: A potential game is a game for which there exists a potential function V: $S \rightarrow$ R such that:

$$\Delta u_i = u_i\left(s_i, s_{-i}\right) - u_i\left(s_i', s_{-i}\right) = \Delta V = V\left(s_i, s_{-i}\right) - V\left(s_i', s_{-i}\right) \qquad \forall i \in P,\ \forall s_i, s_i' \in S_i$$

This definition implies that each player's individual interest is aligned with the groups' interest, since each change in the utility function of each player is directly reflected in the same change for the potential function.

If only one player acts at each time step and that player maximizes or at least improves its utility, given the most recent action of the other players, then the process will always converge to a NE. In addition, global maximization values of the potential function *V* are NE, although they may be just a subset of all NE of the game. These interesting properties of potential games suggest their utilization as an approximation for the optimal value of the network capacity. In this case, the utility function would be equal to the potential function and this equal to the performance metric to maximize. In our case, this performance metric is the global capacity of the network, so the player utility is:

$$u_i\left(s_i, s_{-i}\right) = \sum_{u \in U} C(u) \tag{4.18}$$

Log-Rate Potential Game: The previous games aim at maximizing the global throughput of the network (and therefore, the spectral efficiency). Nevertheless, this kind of solution may lead to solutions where the users experiencing the best channel conditions are allocated more resources than users with poor channel quality. To alleviate this, we propose as well a potential game that aims at maximizing the proportional fairness of the network. With this objective, the player utility is:

$$u_i\left(s_i, s_{-i}\right) = \sum_{u \in U} \log\left(C(u)\right) \tag{4.19}$$

To introduce this fairness in the sharing of the backhaul capacity as well, the allocation of backhaul resources in the backhaul will be the solution of the following optimization problem for each small cell:

$$\begin{aligned} \max \quad & \sum_{u \in S_u} \log\big(C(u)\big) \\ \text{s. t.} \quad & C(u) \le C_{nb}(u) \\ & \sum_{u \in S_u} C(u) \le C_b(S_u) \end{aligned} \tag{4.20}$$

This problem is equivalent to a max-min fairness problem that can be trivially solved.

Timing and decision rules:

Both local and potential games are played as a myopic repeated game. A repeated game is a sequence of stage games where each stage game or step is the same normal form game. The myopic term reflects that each player takes its decision according to the observation of the most recent scenario where it is playing, instead of considering past actions or future expectations. Thus, complex multi-stage strategies are not possible. However, simpler myopic strategies, such as the best or better response dynamics, can be used. A better response is a playing rule that decides to change towards a new strategy that at least improves the current utility. A best response corresponds to a playing rule that decides to change to the strategy that provides the optimum utility given the current opponents' profile. In our case, a better response strategy has been used since generally, its computational complexity is lower than that of the best response strategy.

In addition to the playing rule, a repeated game is characterized by the specific timing followed by the players, that is, the playing order. In this case, a Round Robin scheduling has been used. In Round Robin, at each step only a single player plays. This implies that this player will not update its strategy again until all the remaining players have played. This scheduling guarantees the convergence for the potential game, since any potential game in which players take actions sequentially under a best or better response strategy converges to a pure NE.

4.2.2 Implementation of CT in the iJOIN architecture

Functional Split A) Centralized Connection Control

The proposed algorithms rely on the handover functionality provided in the 3GPP standard. The message sequence chart is shown in Figure 4-11. A Mobility Load Balancing command is triggered by the iSC serving the UE. From this point a similar process to decide if a handover can be performed is executed: first, the UE provides a measurement report which is forwarded to the RANaaS, where the proposed algorithms decide if the UE must change its serving iSC based on the information in the measurement report. If so, the typical handover process is launched.

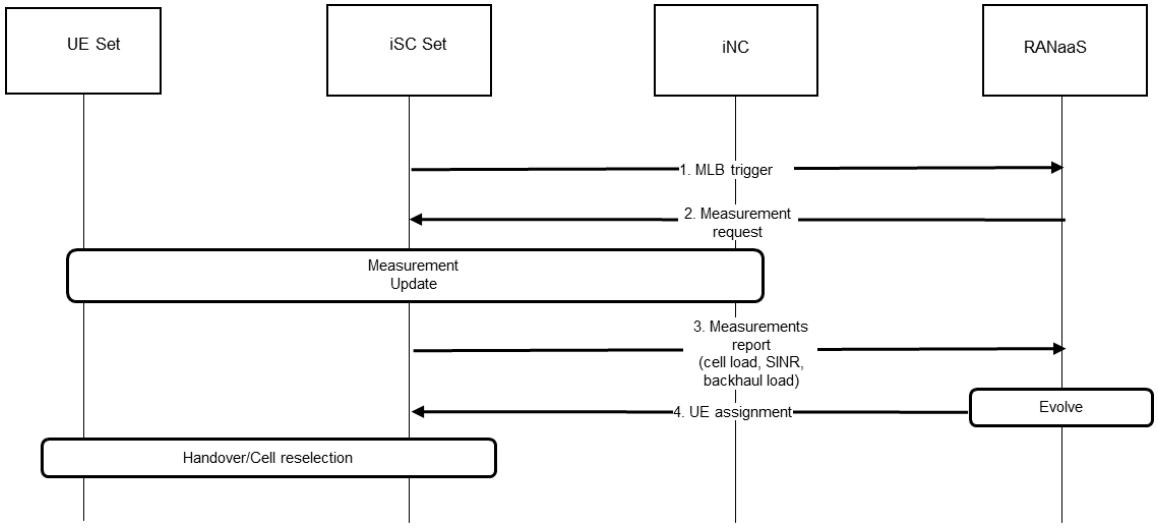


Figure 4-11: Required message passing and functions in the centralized algorithm proposed in CT3.2.

From the system level perspective our algorithm is based on interfaces, functions, and messages that are already standardized, which results in limited complexity. The *Evolve* algorithm can be implemented through

the Mobility Load Balancing (MLB) function, which has been defined in the framework of self-organizing networks (SONs) to improve the LTE performance through coordinated traffic steering [48]. MLB is based on the exchange of information about load level and available capacity amongst neighbouring cells through the X2 interface. Based on these reports, Evolve decides the momentary optimal association amongst UEs and iSCs. According to the output of the algorithm, cell reselection and handover functions are executed to shift idle and connected UEs to the target iSCs. To reduce complexity and system overhead, the periodicity of the reporting can be requested only in the range of 1 to 10 seconds [49]. The process is initiated through the MLB trigger that is sent to the RANaaS by an iSC that is currently overloaded (step 1 in Figure 4-11). Then, the RANaaS requires to the overloaded iSC and to its neighbouring iSCs, measurements on the experienced load, the SINR measured on the radio links, the capacity of the backhaul, etc (step 2). By using the received inputs (step 3), the proposed algorithm can be implemented, and the novel optimal association can be transferred (step 4) to the set of iSCs to be executed.

Functional Split B) Distributed Connection Control

The functional split is very similar to the centralized approach, with the main difference that the distributed algorithm is played in the iSCs themselves. The message sequence chart is shown in Figure 4-11. First, a Mobility Load Balancing command is triggered by the overloaded iSC serving the UE. Then, the UE sends the measurement report to its serving iSC, which distributes it among the surrounding iSCs. Next, the game theoretic algorithm (GAME) is executed based on the information of the measurement report. If the output of the algorithm corresponds to a change in the iSC that serves the UE, a handover process is launched.

As in the centralized algorithm, GAME implementation relies on the MLB function. The information on the load level and the available capacity is exchanged between the neighbouring iSCs that execute GAME to decide on the preferred UE-iSC associations. Any change on the associations is carried out through the cell reselection and handover functions.

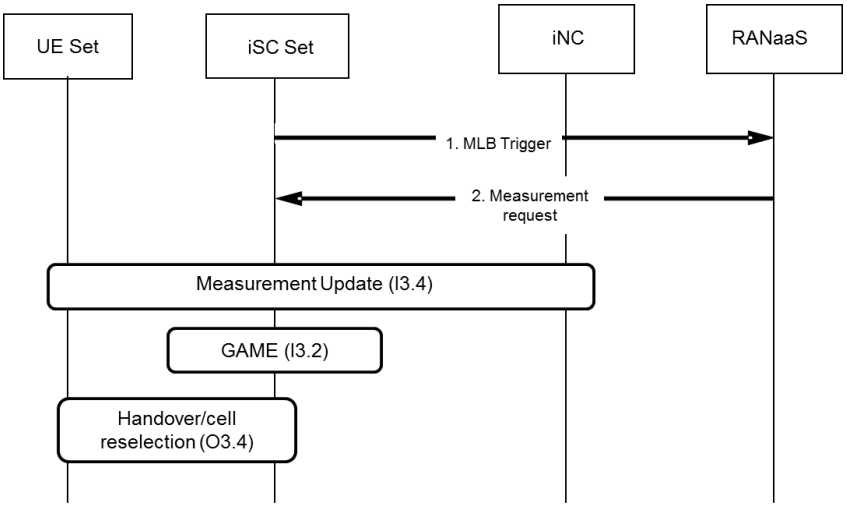


Figure 4-12: Required message passing and functions in the centralized algorithm proposed in CT3.2.

4.2.3 Evaluation of the CT

The performance of the proposed algorithms is evaluated for specific simplified scenarios (considering only one cluster of small cells) using Monte-Carlo simulations. The objective of these simulations is to show the potential benefits of the proposed approach and compare the performance of the different algorithms, identifying the possible trade-offs. The performance of a complete scenario will be evaluated in D3.3.

Compliance with iJOIN objectives

This CT is based on centralized/distributed schemes that attempt to maximize the network capacity by jointly considering the radio access capacity and the backhaul capacity. The proposed solutions perform cell selection algorithms considering not also the radio access but also the capacity available at the backhaul. The

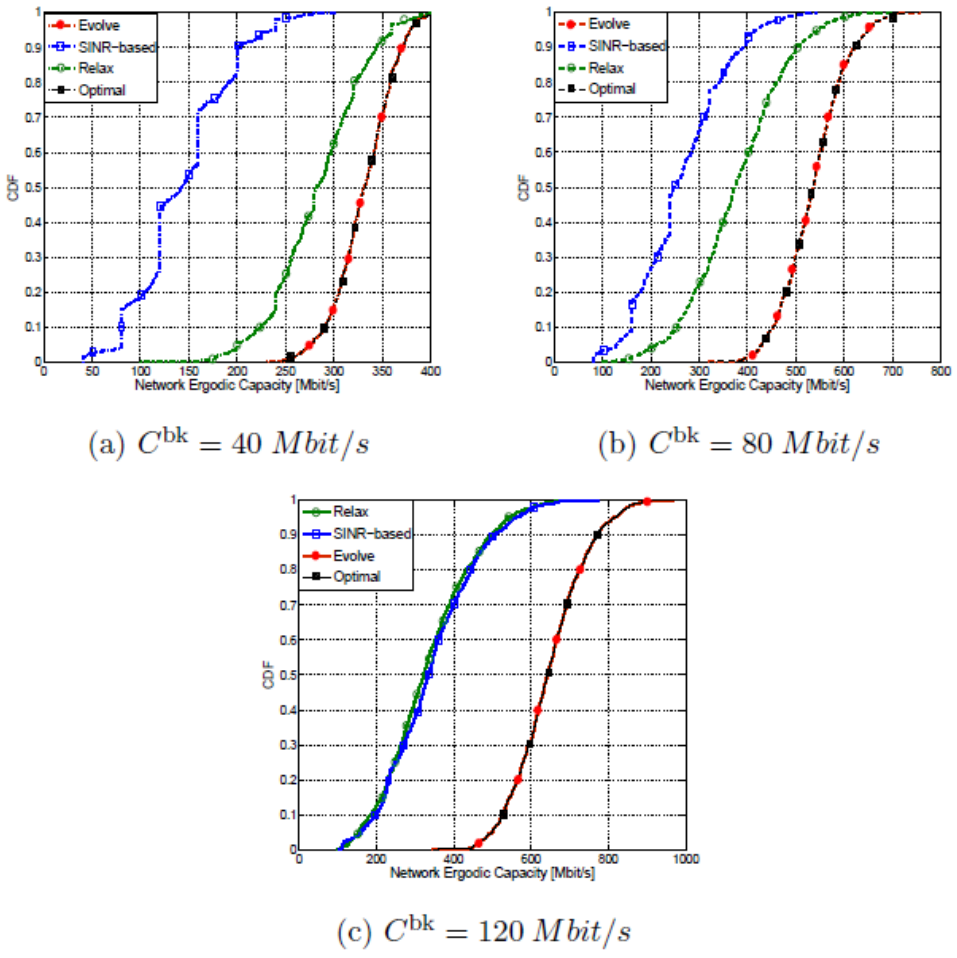
application of these solutions decreases the probability of congestion in low capacity backhaul links, thus increasing the network capacity.

Centralized Connection Control Approach

In this section, we assess the effectiveness of the proposed Evolve algorithm by comparing its performance with respect to the optimal solution, obtained through BF algorithm, and the classical approach where each UE selects the eNB associated with the strongest Reference Signal Received Power (RSRP). We also compare the Evolve algorithm with another connection control scheme proposed in literature, named as Relax [7].

Here we assume that iSCs form 3 × 3 grids located inside the macro cell; moreover, 2/3 of the overall UEs are distributed inside the small cell grids and the remaining UEs are uniformly located in the macro cell area (see Appendix II). Other parameters relevant for this study (such as path loss model and shadowing) follow 3GPP TR 36.872 [21].

The results are averaged over 10^3 independent runs. At the beginning of each run, the clusters of iSCs and UEs are randomly deployed in the macrocell area. In our simulations, UEs include in their active set those eNBs associated with a SINR greater than γ_{th} equals to −3 dB, and the stopping parameter ϵ equals to 0. As already mentioned, we consider full buffer traffic at mobile UEs. Finally, the user SE is upper limited to 12 bit/s/Hz (η_{max}) to fairly evaluate the impact of the RAN and backhaul on the overall network performance.



(a) $C^{\mathrm{bk}} = 40\ Mbit/s$ (b) $C^{\mathrm{bk}} = 80\ Mbit/s$

(c) $C^{\mathrm{bk}} = 120\ Mbit/s$

Figure 4-13: Cumulative distribution function of the Network Shannon Capacity achieved with different association schemes.

Here, we aim to compare the performance of Evolve, Relax, and the SINR-based algorithms with respect to the optimal solution. Due to the high complexity of the BF algorithm, we consider a light deployment scenario, composed by a MeNB, a cluster of 9 iSCs, and 20 UEs located in the central macrocell site. However, additional eNBs located in the surrounding sites are used to model inter-cell interference. Figure 4-13 shows the Cumulative Distribution Function (CDF) of the Network Shannon Capacity C(α) achieved with the MCI scheduling policy with respect to different backhaul constraints. Blue, green, red, and black

lines, respectively, correspond to the SINR-based, the Relax, the Evolve, and the optimal solutions. Moreover, dashed-dotted, dashed, and dotted lines correspond to low, medium, and high backhaul capacity (i.e., C^{bk} equals to 40/80/120 Mbit/s).

Note that in the first case, the backhaul is likely to be the main constraint to the network performance (Figure 5a); hence, the classic SINR-based approach, which only takes into account the quality of the radio link, is characterized by poor performance. However, the higher the backhaul capacity, the lower its impact on the overall capacity: when C^{bk} is set equal to the maximum achievable RAN capacity ($B \cdot \eta_{max}$ =120 Mbit/s), only the quality of the radio links and the network load limit the performance. Therefore, the SINR-based approach achieves more valuable performance, and it gains up to the 133% with respect to the low backhaul capacity case (compare blue lines in Figure 4-13a and Figure 4-13c).

The Relax algorithm enables to improve the performance achieved by the classic SINR-based scheme in low backhaul capacity scenarios, and our simulations show 97% of gain measured at the CDF median value (Figure 4-13a). However, increasing the backhaul capacity the Relax scheme does not result in notable gains any more. This drawback is mainly due to two reasons: first, this approach is based on relaxing the constraint that forces each UE to be served by only an one eNB; however, this may lead to a solution that diverges from the optimal one.

Second, the Relax scheme does not guarantee to improve the network capacity during its iterative process. On the contrary, we note as Evolve has the same performance as the optimal solution (red and black lines are superposed) and gains up to 132% and 100% with respect to the classic SINR-based and Relax schemes (measured at the median value of Figure 4-13a and Figure 4-13c, respectively). The gain with respect to the SINR-based algorithm is due to the load- and backhaul-aware properties of the proposed scheme that better balances service requests across the network and increases the overall resource utilization.

Distributed Connection Control Approach

Description of the baseline used for the evaluation of the CT

For benchmarking, we compare the network capacity obtained with the proposed algorithm to the one that would be obtained if the users were always associated to the small cell received with the strongest SINR, since this is the most common procedure to perform cell selection in wireless networks.

Discussion of results of the CT

The evaluation is carried out in a scenario formed by a cluster of 10 small cells and 40 users. The backhaul capacity of each small cell is a uniform random variable ranging from 20 to 40 Mbps. The power of each small cell is 41 dBm and the small cells are randomly deployed in a circular area of radius 50m.

First, we compare the CDF of the capacity of proposed algorithm with that of the benchmark (user always associates to the small cell with strongest signal). As can been in Figure 4-14, the proposed solution improves the total capacity of the network, being the average increase in capacity of about a 5%. In general, this value will depend on the diversity of possible backhaul capacities for the small cells.

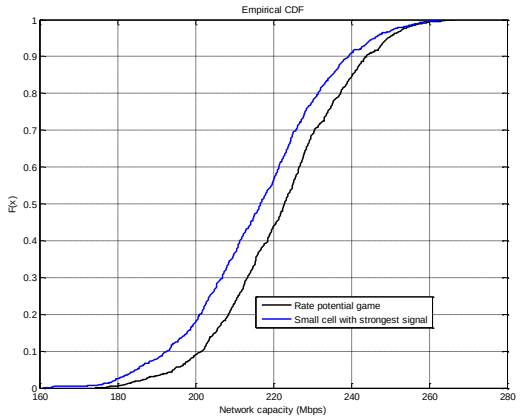


Figure 4-14 CDF of the proposed solution vs benchmark

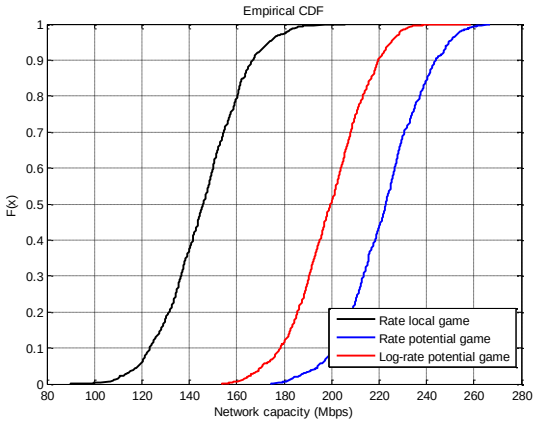


Figure 4-15 Comparison of the network capacity achieved with the different games

Figure 4-15 compares the different network capacity obtained with the different proposed games. As expected, the rate local game is the worst one since in this case the utility function is not aligned with the performance metric (the network capacity). Nevertheless, its computational complexity is the lowest one. Additionally, the rate potential game also outperforms the network capacity of the log-rate potential game. Nevertheless, if we compare the CDF of the log-sum rate of the network (which is a measure of the fairness of the network) with that obtained with the rate potential game (Figure 4-16), the last one obtains the best results. This result shows the inevitable trade-off between the maximization of the network capacity (and also the spectral efficiency) and the degree of fairness that is achieved in the network when the transmission resources are shared amongst the users.

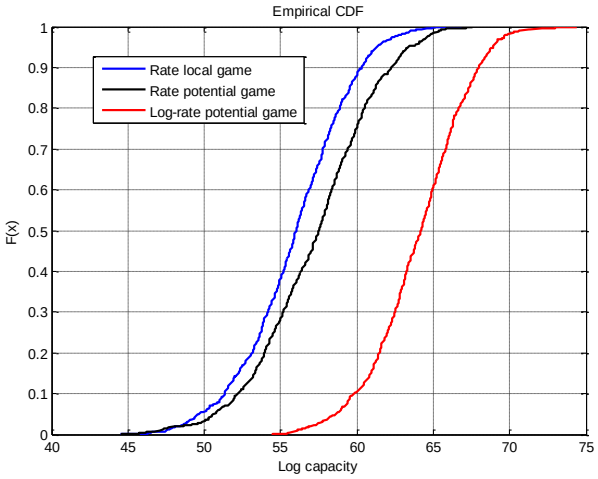


Figure 4-16: Comparison of the network log-sum rate achieved with the different games

4.3 CT 3.3: Energy-Efficient MAC/RRM at Access and Backhaul

4.3.1 Technical description

Scenario

We consider a scenario where the RANaaS manages the energy activity of *N* iSCs according to the network access characteristics and the QoS constraints. The main objective is to find an optimal policy to centrally decide the patterns associated to the discontinuous transmission (DTX) mode at iSCs. Markov decision process is used to model this optimization problem and to find the associated solution [8].

System Model

In this CT, the RANaaS receives data from the core network through the S1 interface and stores it in a dedicated buffer. When required, the RANaaS, activates a given iSC and forwards to it the associated traffic through the J1 interface. Thereafter, the activated iSC will autonomously manage available radio resources to

efficiently transmit the data received from the RANaaS according to a first-in-first-out policy. In our model the RANaaS is equipped with *N* buffers of size *M* packets, each one dedicated to a specific iSC.

These buffers are continuously supplied by new data; here we indicate as $f_{j,t}$ the number of packets received at the buffer $j \in [1;N]$ at the time step *t*, where each element $f_{j,t}$ belongs to a finite set {0,….,F}. Hence, we can define the status of a buffer *j* as $\boldsymbol{q}_{j,t}=(\mathbf{L}_{j,t};\ n_{j,t}) \in Q$, where $n_{j,t} \in \{0,\ldots,M\}$ is the number of packets present in the queue and $\mathbf{L}_{j,t}$ is a length $n_{j,t}$ vector with entries $l_{i,j,t} \in \{0,\ldots,L\}$ that maintains the time-to-live (TTL) of the packets in the queue. When activated, an iSC *j* transmits at most $r_{j,t} \in \{0,\ldots,R\}$ packets, which depends on the available bandwidth and the spectral efficiency of the selected modulation and coding scheme.

Let *S* be a finite set referred to as the network state space and defined as $S = Q \times R \times F$, where *Q*, *R*, and *F* are the composite state spaces, which describe the buffer state, the data rate, and the incoming traffic of the iSCs. At each time step, the RANaaS observes the current state of the network $s_t \in S$ and selects an action $\mathbf{a_t}$ from the set *A*, $\{\mathbf{a_0}, \mathbf{a_1},\ldots,\mathbf{a_N}\}$, where $\mathbf{a_0}$ is a length N null vector and $\mathbf{a_j}$, (0,…,1,….,0) is a length N vector with entry *j* equals to 1 and all other entries equal to 0. In particular, $\mathbf{a_0}$ represents the action that maintains all the iSCs idle and the other actions correspond to activate one amongst the deployed iSCs and keep the others idle, i.e., $\sum_{j=1}^{N} a_{j,t} \le 1 \ \forall\ \mathbf{a} \in A$. The implementation of the decided action changes the current status from s_t to s_{t+1}, with a state transition probability $T(s_t \mid s_{t+1};\ a_t)$, and it incurs in an immediate cost $C_t = C(s_t, a_t)$.

Our goal is to find an optimal policy π^* that associates an action $a_t(s_t \mid \pi^*)$ to the state s_t. This policy has to manage the activity of the iSCs in order to minimize the energy consumption and satisfy the QoS constraints while avoiding simultaneous access of multiple iSCs to the same frequency resources in an uncoordinated manner.

A. Assumptions on the data rate and traffic dynamics:

I. The achievable data rate at each iSC evolves according to an ergodic Markov chain with transition probabilities $T(r_{j,t+1} \mid r_{j,t})$ independent of the time step, action, buffer state, and incoming traffic.

II. The incoming traffic at each iSC is modelled as an ergodic Markov chain with transition probabilities $T(f_{j,t+1} \mid f_{j,t})$ independent of the time step, action, buffer state, and data rate.

In our model the buffer occupancy evolves according to the Lindley's recursion [9], which we have opportunely modified to take into account the latency constraints:

$$n_{j,t+1} = \min\left\{\left[n_{j,t} - r_{j,t}\cdot a_{j,t}\right]^+ - d^l_{j,t} + f_{j,t+1}, M\right\}, \tag{4.21}$$

where $[x]^+$ returns x if $x > 0$ and 0 otherwise. Furthermore, the number of bits dropped due to latency constraints $d^l_{j,t}$ can be computed as

$$d^l_{j,t} = \left[\sum_{i=1}^{n_{j,t}} \delta(l_{j,i,t} - 1) - r_{j,t}\cdot a_{j,t}\right]^+, \tag{4.22}$$

and the number of bits dropped because of the limited buffer size is

$$d^b_{j,t+1} = \left[\left[n_{j,t} - r_{j,t}\cdot a_{j,t}\right]^+ - d^l_{j,t} + f_{j,t+1} - M\right]^+. \tag{4.23}$$

The overall number of bits lost at each time step is equal to the sum of the number of bits dropped because of unsatisfied latency constraints and the number of bits lost for the limited buffer size

$$d_{j,t} = d^l_{j,t} + d^b_{j,t}. \tag{4.24}$$

Finally, the TTL of the packets located in the queues evolves according the following rule:

$$\forall\ j \in [1; N], \forall\ i \in [1; n_{j,t+1}]$$

$$l_{j,i,t+1} = \begin{cases} l_{j,i+r_{j,t}\cdot a_{j,t}+d^{l}{}_{j,t}+d^{b}_{j,t+1},t} - 1, & i \le \left[n_{j,t} - r_{j,t}\cdot a_{j,t}\right]^{+} - d^{l}_{j,t} - d^{b}_{j,t+1} \\ L, & otherwise \end{cases} \quad (4.25)$$

Our optimization problem can be represented as follows:

- a set of states *S*;
- a set of actions *A*;
- a state transition probability *T*;
- a cost function *C*.

The model is Markovian since the state transition depends only on the current status and is independent of any previous environment states or agent actions. Accordingly, the transition probabilities between the current state s_t and the next state s_{t+1} can be modelled as

$$T(s_{t+1} / s_t, a_t) = \prod_{j=1}^{N} T(r_{j,t+1} / r_{j,t}) \cdot T(f_{j,t+1} / f_{j,t}) \cdot I_{n_{j,t+1}=n_{j,t+1}(n_{j,t},r_{j,t},a_{j,t},d^{l}_{j,t},f_{j,t+1},M)} \cdot I_{L_{j,t+1}=L_{j,t+1}(L_{j,t},n_{j,t},r_{j,t},a_{j,t},d^{l}_{j,t},d^{b}_{j,t+1})}, \quad (4.26)$$

where I_x is the indicator function that returns 1 if *x* is true and 0 otherwise, and the functions $n_{j,t+1}(\cdot)$ and $L_{j,t+1}(\cdot)$ are defined in (4.21) and (4.25) , respectively. Note that the dropped bits are not part of the system state but they affect the buffer state evolution.

Since we are looking for an optimal policy in the sense of the energy efficiency and system QoS, we can define the system cost function as

$$C_t = \tilde{P}(s_t, a_t) + \alpha \cdot \tilde{d}(s_t, a_t), \quad (4.27)$$

where α is a weight factor that prioritizes between energy efficiency and QoS and $\tilde{P}(s_t, a_t)$ and $\tilde{d}(s_t, a_t)$ are the momentary sum of the N length vectors that indicate the power consumption at the iSCs and the dropped packets.

Approach

When we consider a classic Markov Decision Process (MDP), there exist dynamic programming algorithms (such as the value iteration [8]) that enable to find the optimal deterministic stationary policy π^* that minimizes the total expected cost, which is usually discounted by a factor $\gamma \in [0; 1)$ in case of infinite time horizon [10]. Hence, the optimal value of the state *s* is defined as the expected discounted cost if the system starts at the state $s \in S$ and follow the policy π^*_γ:

$$V^*_\gamma(s) = \min_{\pi} E\left\{\sum_{t=0}^{\infty} \gamma^t \cdot C_t\right\} \quad (4.28)$$

According to the Bellman's principle [11], the above optimal value is unique and can be found solving the following equation

$$V^*_\gamma(s) = \min_{a \in A}\left(\overline{C(s,a)} + \gamma \sum_{s' \in S} T(s'/s,a) \cdot V^*_\gamma(s')\right) \quad (4.29)$$

which states that the optimal value of *s* is the expected cost $\overline{C(s,a)} = E\{C(s,a)\}$ plus the expected discount value of the next state *s'* when using the optimal action. Hence, the optimal policy π^*_γ can be defined as

$$\pi_{\gamma}^{*} = \arg\min_{a \in A} \left(\overline{C(s,a)} + \gamma \sum_{s' \in S} T(s'/s,a) \cdot V_{\gamma}^{*}(s') \right) \tag{4.30}$$

In our problem, the cost function includes two components that determine an optimization trade-off; to limit packet drop, the system requires continuously activating the small cells and transmitting the packets stored in the buffer. On the other hand, minimizing the energy consumption needs to maintain the small cells idle as long as possible, which in turns results in packet loss.

In general, since the optimal weighting factor α^{*} in (4.27) is unknown, we cannot identify a-priori a single optimal policy, but we have to compute the set of deterministic stationary Pareto efficient policies, and then select one of them [12]. However, by transforming our problem in a constrained MDP, we can find a straightforward method to find an optimal stationary policy.

Let $\pi_{\gamma,\alpha}^{*}$ be the optimal policy that minimizes $V_{\gamma,\alpha}(s) = \tilde{P}_{\gamma,\alpha}(s) + \alpha \cdot \tilde{d}_{\gamma,\alpha}(s)$, where $\tilde{P}_{\gamma,\alpha}(s)$ and $\tilde{d}_{\gamma,\alpha}(s)$ are the expected discounted power consumption and dropped packet, respectively.

Hence, the following statement holds [13]:

Theorem: $\alpha \rightarrow \tilde{d}_{\gamma,\alpha}^{*}(s)$ is not increasing.

Corollary: Let $d_{\max}$ be the maximum admissible number of packet dropped in the small cell network, and assume that exists α^{*} s.t. $\tilde{d}_{\gamma,\alpha^{*}}^{*}(s) = d_{\max}$. Therefore, $\pi_{\gamma,\alpha^{*}}^{*}$ is the policy that minimizes $\tilde{P}_{\gamma,\alpha}(s)$ subject to $\tilde{d}_{\gamma,\alpha}^{*}(s) \leq d_{\max}$.

Therefore, to find the optimal small cell activation controller, we may start from a given value of α and implement the value iteration algorithm. While $\tilde{d}_{\gamma,\alpha}^{*}(s) > d_{\max}$, we keep decreasing α until the number of dropped packets satisfies the system constraint, then we have the optimal controller.

4.3.2 Implementation of CT in the iJOIN architecture

In this CT, the RANaaS receives data from the core network through the S1 interface. Therefore, when QoS constraints discriminate some packets as urgent, the RANaaS activates the corresponding iSC (1) and commands to perform CSI estimation (2). According to the received measurements (3) as well as past information, the RANaaS forwards to the iSC the amount of data that can be transmitted to the UEs associated to the activate iSC (4). After the data transmission process has been finalized, the iSC can be de-activated.

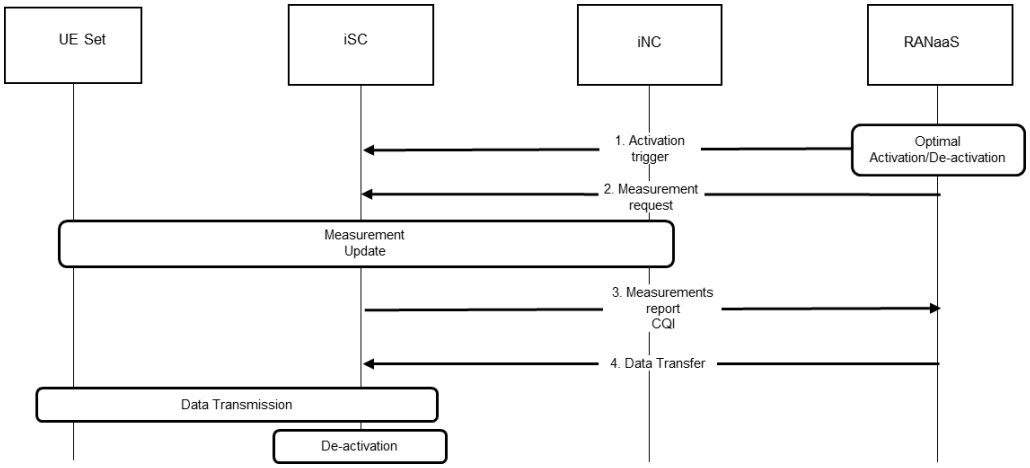


Figure 4-17: Message sequence chart for CT 3.3

4.3.3 Evaluation of the CT

Compliance with iJOIN objectives

This CT aims to reduce the energy consumption associated to the iJOIN radio access network without affecting the QoS perceived at the end users.

Description of the baseline used for the evaluation

We will compare the proposed solution with more simple approaches, i.e., random activation and greedy (i.e., myopic) mechanisms in terms of energy efficiency and packet losses.

Discussion of results of the CT

We will evaluate the energy consumption of the proposed framework by extending power consumption models already available in literature, which consider a cloud platform, the backhaul, and base stations [58].

To provide performance evaluation, we consider a hotspot composed by four small cells (N=4), which are coordinated by the nearby RANaaS. Without loss of generality, we consider that in a given transmission time interval (set equal to 1ms), at most 1 packet is received at each buffer and that activated small cells can simultaneously transmit up to two packets. Packet length is considered fixed and equals to 1Kbits. Moreover, we assume that small cell and backhaul equipment in idle mode consume 60% of their zero load power. Finally, we consider that 5% of the base-band signal processing load is transferred by the small cell to the RANaaS to manage their activation/deactivation.

Figure 4-18 illustrates the performance of the optimal stationary solution (obtained through value iteration [8]) with respect to a random small cell management policy and greedy approach, where action is taken to minimize the instantaneous value of the cost function, i.e., without taking into account the total (over time) expected cost. The red solid line, black dotted line, and blue dashed line respectively correspond to the optimal, greedy, and random solution. Performance are presented in terms of cumulative network energy efficiency [bit/W], which is computed as the ratio of the cumulative number of transmitted bits and the associated cumulative network energy consumption measured over 1 second. The optimal solution leads up to 96% of gain with respect to the greedy solution; however, it gains only 16.4% with respect to the random policy. This surprising result is due to limited energy consumption associated with the random solution, which in turns results in an unacceptable number of dropped data.

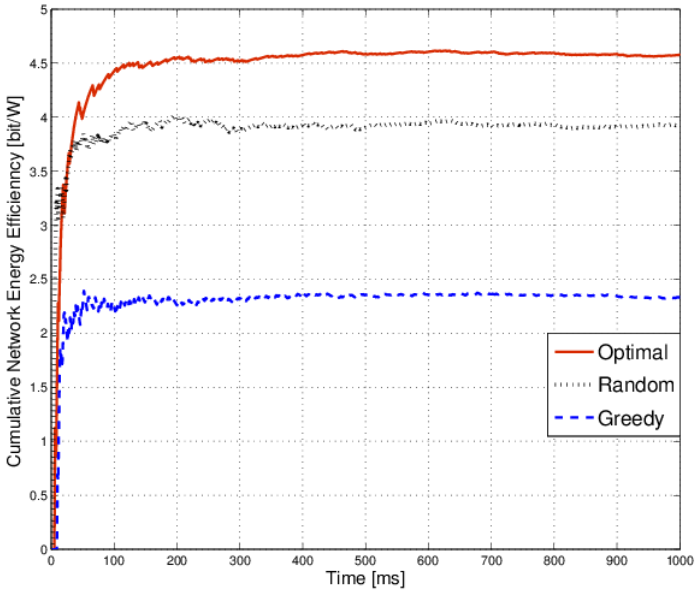


Figure 4-18: Cumulative Network EE with respect to different small cell management scheme.

4.4 CT 3.4: Computational Complexity and Semi-Deterministic Scheduling

4.4.1 Technical description

Scenario

The computational resources of cloud platforms enable centralized processing of complex tasks with global knowledge, which is not available at the individual base stations. Semi-deterministic scheduling exploits these resources by shifting the computational load partially into the cloud, thus enabling the creation of a global scheduling plan for very dense small cell deployments. This is necessary to combat the severe inter-cell interference caused by short inter-site distances in such scenarios.

The challenges for semi-deterministic scheduling are two-fold:

- First, to identify the maximum achievable performance, considering constraints on computational resources and backhaul. For example, if the backhaul delay is high, the channel may change significantly before the channel information arrives at the central processor (CP). Therefore, the computation needs to be based on averaged or compressed information, leading to more coarse/long-term scheduling plan.
- Second, to develop actual multi-level scheduling algorithms to exploit centralization gains in cloud-processing considering backhaul constraints at the network edges. Here, the challenge to identify the minimum amount of signalling which is required to pass the channel state information to the CP, to determine the optimal schedule, and to provide the scheduling decision to the individual base stations arises.

As depicted in Figure 4-19, the scheduler is divided into two stages: a coarse-grain scheduler at the central entity (the RANaaS) based on global but imperfect (quantized and outdated) channel state information, and a second stage at the iSC based on local but less quantized and outdated channel state information (CSI).

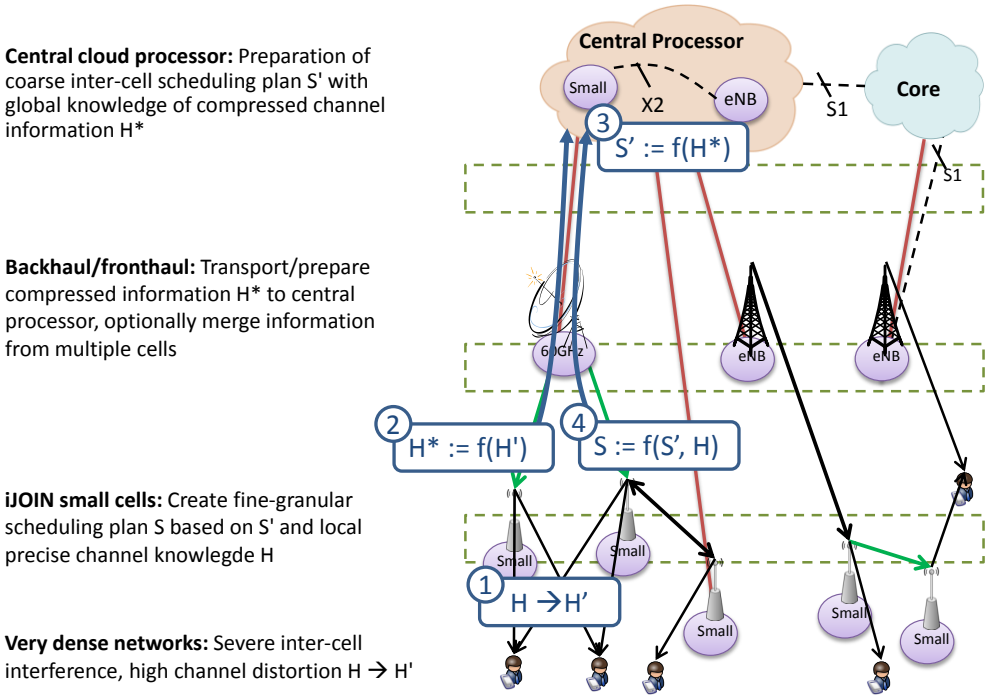


Figure 4-19: Semi-deterministic, hierarchical scheduling.

System Model

For scheduling the users, only imperfect channel knowledge is available at the CP as well as at iSCs. Due to the backhaul limitations in terms of capacity and/or latency, the channel uncertainty at the CP is larger than the uncertainty at the iSC. In order to model imperfect channel knowledge, the actual channel of user *u* can be expressed as the sum of the available channel estimate and a zero mean error which reflects the uncertainty of the channel [24], as

$$h_u = \hat{h}_u + e_u. \tag{4.31}$$

Note that the model is based on the assumption of minimum mean square error (MMSE) estimation, which inherently lowers the power of the estimate with an increasing error variance. The sources of CSI impairments considered in this CT are: the channel estimation errors, the feedback quantization and the delays between channel observation and transmission. Since the channel is assumed to be Gaussian distributed,

$$h_u \sim \mathcal{N}_{\mathrm{C}}(0, \lambda_u), \tag{4.32}$$

the error model $e_u \sim \mathcal{N}_{\mathrm{C}}(0, \lambda_u \varepsilon_u)$ can also be assumed to be Gaussian. Consequently, the power of the channel estimate $\mathbf{E}\{|\hat{h}_u|^2\} = \lambda_u (1-\varepsilon_u)$ decreases with the error variance ε_u.

Approach

In this paragraph, we mathematically derive the expression for obtaining the rate which needs to be assigned for transmission in order to guarantee a certain outage probability with imperfect CSI. In the following, the amplitude of the channel and its estimate is written as $g_u = |h_u|$ and $\hat{g}_u = |\hat{h}_u|$, respectively. Furthermore, the user index *u* is omitted to improve readability.

Based on Shannon's formula the maximum rate which is achievable at an SNR $\rho\lambda / \sigma_n^2$ is

$$R = \log_2\left(1 + \rho g^2\right), \tag{4.33}$$

where ρ is the transmit power at the iSC and the receiver noise $\sigma_n^2 = 1$ is normalized to one. In order to achieve the rate supported by the current channel state, the rate in (4.33) needs to be assigned for transmission. Since (4.33) includes the current channel state, the rate which is actually supported by the channel is not perfectly known by the iSC. Consequently, in this setup two basic cases can occur. Either the iSC allocates a rate $\breve{R}$ which is equal or below the actual one (transmission is successful), or outage occurs in case the allocated rate exceeds the one supported by the channel. Hence, the probability of outage results in:

$$p_{\text{out}} = \mathbf{P}\left\{\log_2\left(1 + \rho g^2\right) < \breve{R} | \hat{g}\right\}. \tag{4.34}$$

By rearranging (4.34), the outage probability equals the cumulative distribution function (CDF) of the actual known channel at the point $\sqrt{(2^{\breve{R}} - 1)/\rho}$

$$\begin{aligned} p_{\text{out}} &= \mathbf{P}\left\{g < \sqrt{(2^{\breve{R}} - 1)/\rho} | \hat{g}\right\} \\ &= F_{g|\hat{g}}\left(\sqrt{(2^{\breve{R}} - 1)/\rho}\right). \end{aligned} \tag{4.35}$$

Since the channel known at the transmitter side is a complex Gaussian distributed random variable with mean $\hat{h}$, the probability density function (pdf) of its amplitude follows a Rician distribution, as

$$f_{g|\hat{g}}(g) = \frac{2g}{\lambda\varepsilon} \exp\left(-\frac{g^2 + \hat{g}^2}{\lambda\varepsilon}\right) I_0\left(\frac{2g\hat{g}}{\lambda\varepsilon}\right), \tag{4.36}$$

where $I_0(x) = \sum_{l=0}^{\infty} (x/2)^{2l} / (\Gamma(l+1) l!)$ is the modified Bessel function of the first kind and order zero and Γ is the Gamma function. Integrating over (4.36) gives the CDF of the known channel

$$\begin{aligned} F_{g|\hat{g}}(b) &= \int_0^b f_{g|\hat{g}}(g) \\ &= 1 - Q_1\left(\sqrt{\frac{2\hat{g}^2}{\lambda\varepsilon}}, \sqrt{\frac{2b^2}{\lambda\varepsilon}}\right) \\ &= 1 - \exp\left(-\frac{\hat{g}^2 + b^2}{\lambda\varepsilon}\right) \sum_{m=0}^{\infty} \left(\frac{\hat{g}}{g}\right)^m I_m\left(\frac{2\hat{g}b}{\lambda\varepsilon}\right), \end{aligned} \tag{4.37}$$

where Q_1 is the Marcum Q-function and $I_m(x) = \sum_{l=0}^{\infty} (x/2)^{2l+m} / (\Gamma(l+m+1)l!)$ is the modified Bessel function of the first kind and order *m*.

4.4.2 Implementation of CT in the iJOIN architecture

The proposed scheduling algorithm is divided into multiple stages (see Figure 4-20). The local schedulers at the iSCs combine recent CSI obtained from their UEs with the global scheduling selection made at the RANaaS. Although the global scheduler performs its algorithm based on more outdated CSI (due to backhaul latencies), it receives channel information from all iSCs. Consequently, the RANaaS selects a PRB allocation for all iSCs, which are dedicated to the respective veNB. However, each iSC has the opportunity to set aside the global selection and reallocate resources if respective channel states change with a certain amount.

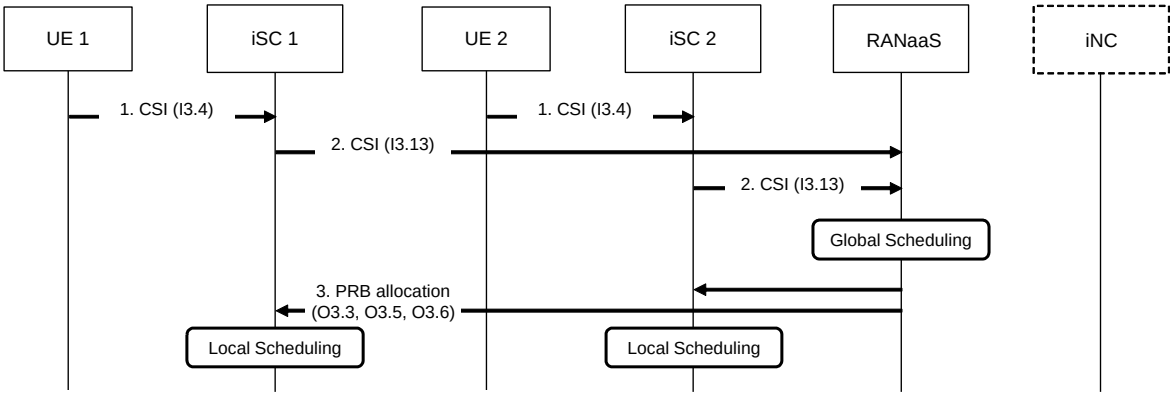


Figure 4-20: Message sequence chart for semi-deterministic scheduling in CT3.4.

The preferred functional split is "coordinated RRC," where the MAC layer is implemented decentralized at the iSCs and the MAC layer is shifted towards the RANaaS.

4.4.3 Evaluation of the CT

The performance of the developed algorithms is first evaluated analytically as well as by means of Monte-Carlo simulations for specific toy examples. Those analyses are intended to show the basic behaviour and certain effects of the algorithm. Moreover, the performance within a larger network will be evaluated via system level simulations at a later stage.

Compliance with iJOIN objectives

This CT aims to increase the throughput of the system, while long term fairness is considered by utilizing proportional fair scheduling. The algorithm takes into account that channel knowledge is only imperfectly available at the base stations. Consequently, the rate adaptation might allocate rates, which are not supported by the underlying radio channel, leading to outages. The algorithm shall ensure to reach a certain outage probability regarding a longer time frame.

Description of the baseline used for the evaluation of the CT

For benchmarking, the scheduling algorithm is compared with a scheduler which ignores the effect of imperfect channel information and treats the available knowledge as perfect. Therefore the proportional fair scheduler is used as given in [25] for the multi base station case.

Discussion of results of the CT

The evaluation given in this document captures basic relations in order to give insight to the underlying mechanisms, relevant for deriving a proportional fair scheduling algorithm which considers CSI imperfections and makes use of the knowledge of the error variance. The given results are obtained analytically as well as by Monte-Carlo simulations over several channel realizations or realizations of the channels estimate.

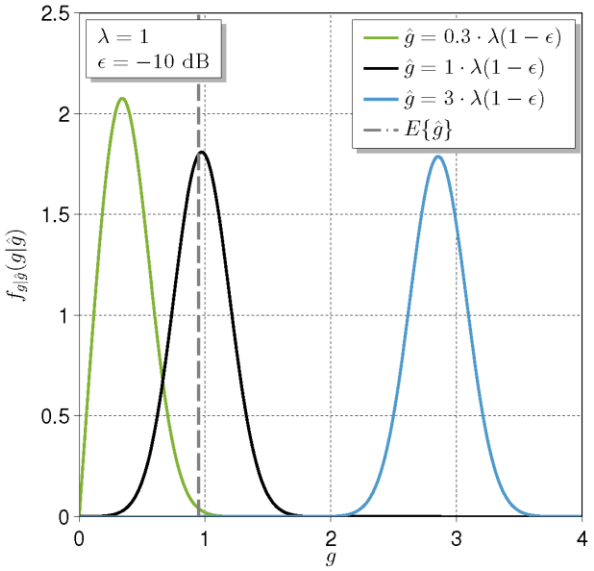


Figure 4-21: PDF of the known channel for different amplitudes

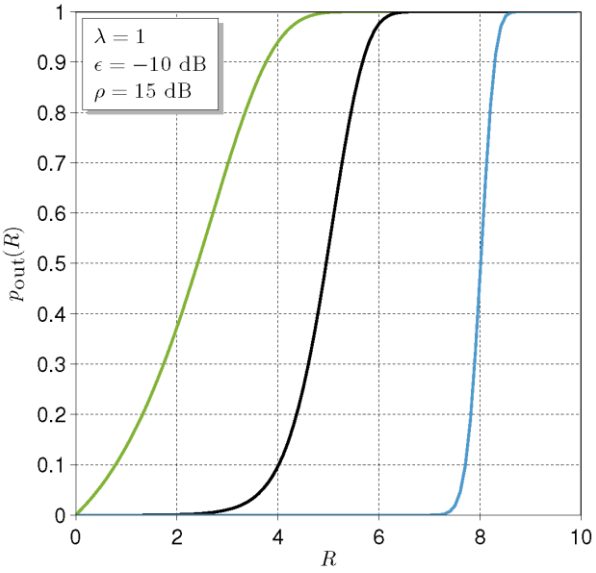


Figure 4-22: Outage probability as a function of the allocated rate

Figure 4-21 illustrates the pdf of the channel's amplitude as it is known at the iSC or the CP, for three different estimated amplitudes (green, black and blue curve). The corresponding outage probabilities as a function of the allocated rates are plotted in Figure 4-22 for an SNR of 15 dB.

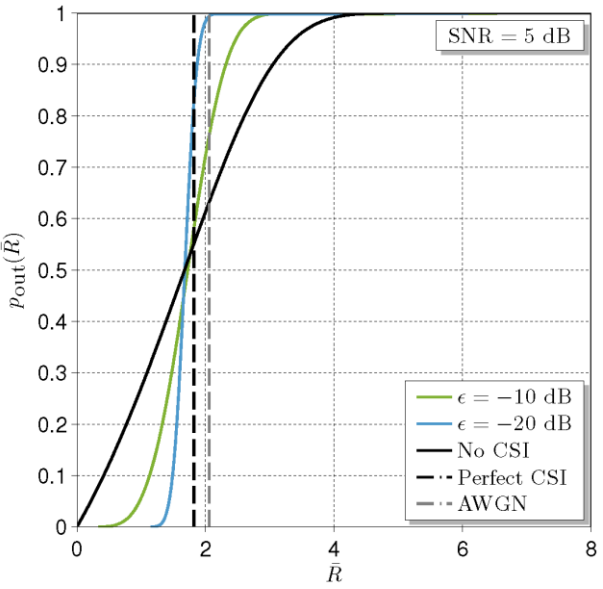


Figure 4-23: Outage probability as a function of the allocated rate amplitudes for an SNR of 5 dB.

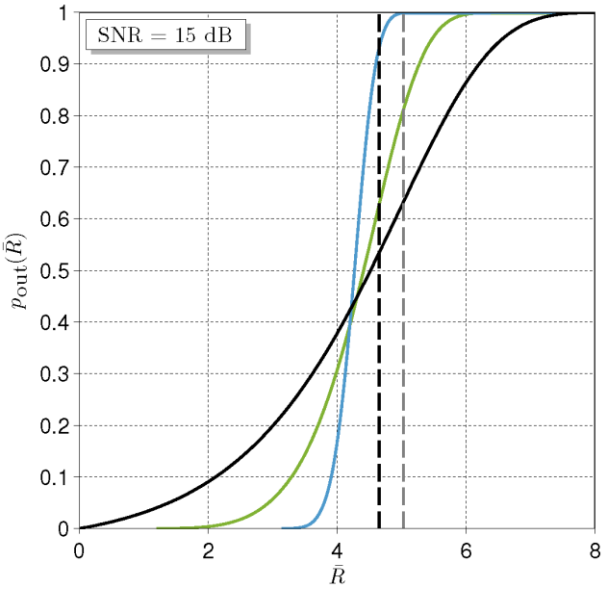


Figure 4-24: Outage probability as a function of the allocated rate for an SNR of 15 dB.

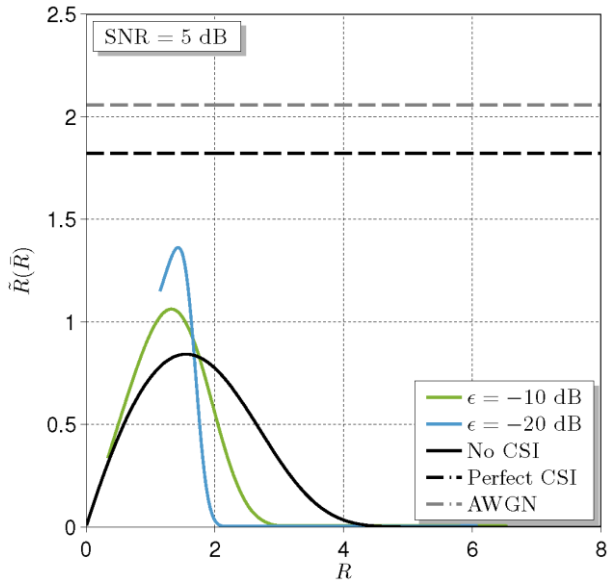


Figure 4-25: Net rate as a function of the allocated rate for an SNR of 5 dB.

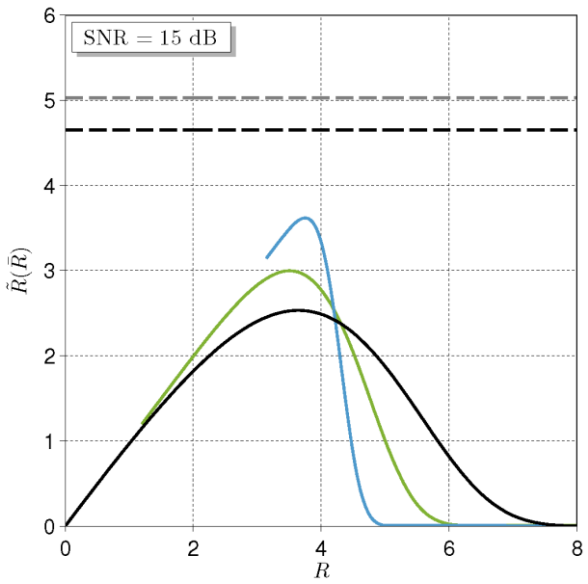


Figure 4-26: Net rate as a function of the allocated rate for an SNR of 15 dB.

While Figure 4-22 shows the outage probability for given channel estimates, the ergodic results achieved by averaging over the actual available channel estimate is given in Figure 4-23 and Figure 4-24 for an SNR of 5 dB and 15 dB, respectively. It can be observed that in both cases for achieving a certain outage probability, the rate which needs to be allocated decreases with the error variance (at least for outage probabilities below 0.4). Current results presented in Figure 4-23 and Figure 4-24 show the basic principle for a scheduling algorithm, to be developed in the next phase of iJOIN. The plots illustrate the impact of the CSI impairment to the rate which shall be allocated on average in order to achieve a certain outage probability. A predefined outage probability is especially of interest if certain delay requirements need to be hold. Another option is to use the outage probability which maximizes the net rate of a user, as illustrated in Figure 4-25 and Figure 4-26 for an SNR of 5 dB and 15 dB, respectively. It can be observed, that the rate which need to be allocated to achieve the optimum is not a monotonic function of the error variance.

4.5 CT 3.5: Cooperative RRM for Inter-Cell Interference Coordination in RANaaS

4.5.1 Technical description

Scenario

We consider a dense deployment of iSCs, which all operate on the same frequency to improve the spatial reuse. Such a technical solution leads to high co-channel interference, which can significantly degrade the potential performance gain of small cells. iSCs are inter-connected through a backhaul, which is characterized by limited capacity and finite latency.

Cooperation enables the implementation of ICIC mechanisms, which can improve transmission robustness and maximize the network capacity. Moreover, the cellular network can exploit the iJOIN RANaaS architecture to flexibly implement ICIC functionalities either in a centralized or a distributed fashion.

System Model

The system is considered as a multi-cell LTE network that consists of a dense deployment of small cells. For our study, we consider the downlink only. The small cell network consists of L iSCs. Each iSC serves M_l users and the total number of users in the system is the aggregation of the users of all L iSCs, such that $M_T = \sum_{l \in L} M_l$.

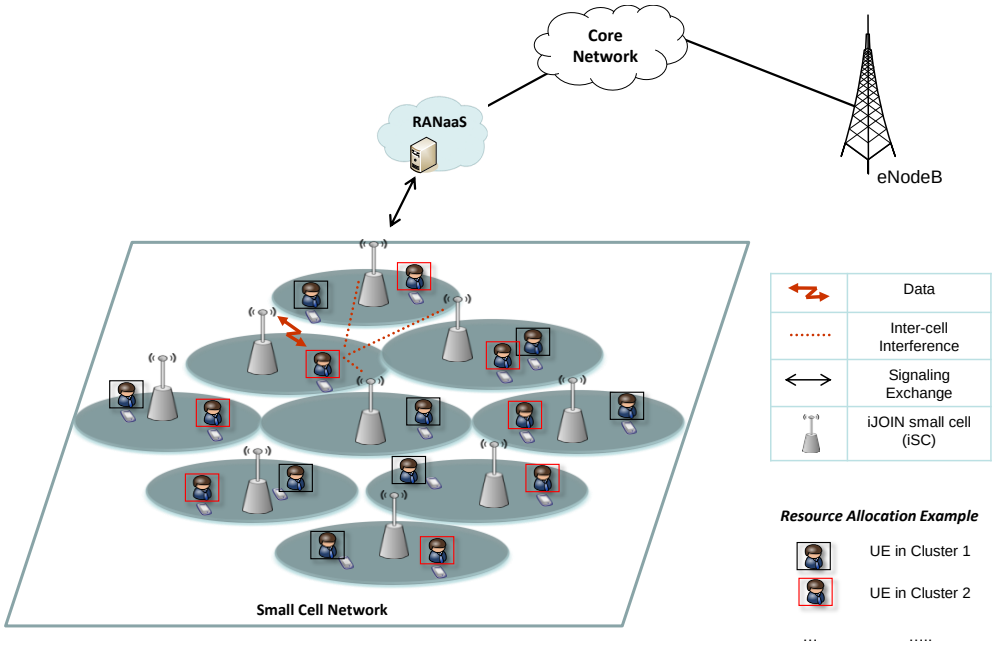


Figure 4-27: Inter-cell interference coordination between iSC

Here, $m(l) \in M_l$ represents the user attached to iSC l, for $L = \{l \mid \forall l \in 1,2,..,L\}$ assuming each user is served by only one iSC. This system also includes a central entity (i.e. RANaaS) that acts as the control unit that resolves the conflicts (in terms of interference) in the small cell network.

In the small cell network, the problem of network optimization can be translated to a weighted sum rate maximization problem, where the weight factors can be tuned accordingly to maintain fairness or other per user service requirements of the network. Let $\{w_{m,n}, m \in M_T, n \in N\}$ be arbitrary user weights taking into account instantaneous QoS requirements and $R_{m(l),n}$ the achievable user's data rate in terms of spectral efficiency on each sub-channel (using the truncated Shannon capacity formula) and is represented as: $R_{m(l),n} = \log_2(1+\rho SINR_{m(l),n})$, where ρ accounts for the SNR gap observed in practice in a system using adaptive modulation and coding. A useful approximation of ρ can be given as $\rho = -1.5/\ln(5 \cdot BER)$ that assumes QAM detection for a given Bit Error Rate (BER). The corresponding SINR is $SINR_{m(l),n} = P_{l,n} G_{m(l),l,n} \Big/ (\sum_{i \neq l \in I_{m(l),n}} P_{i,n} G_{m(l),i,n} + \eta)$, where $P_{l,n}$ is the small cell transmit power and

$G_{m(l),l,n}$ is the channel gain between iSC *l* and UE *m* in the sub-channel *n*. Moreover, η is the power of the thermal noise and $I_{m(l),n}$ accounts for the set of the interferers in a specific sub-channel *n*.

The optimization problem is to find the optimal resource allocation (subcarrier and power control) in order to maximize the weighted sum-rate:

$$\max_{A,P} \sum_{l=1}^{L} \sum_{m(l)\in M_l} \sum_{n\in N} w_{m(l),n} R_{m(l),n} a_{m(l),n} \tag{4.38}$$

Subject to:

$$a_{m(l),n} \in \{0,1\}, \forall l \in L, n \in N \tag{4.39}$$

$$\sum_{n=1}^{N} P_{l,n} \le P_{l,\max} \tag{4.40}$$

$$\sum_{m(l)\in M_l} a_{m(l),\mathrm{n}} \le 1, \forall l \in L, n \in N \tag{4.41}$$

where $A = \{a_{m(l),n} \mid a_{m(l),n} \in \{0,1\}\}$ is the binary variable corresponding to the allocation decision for the sub-channel *n* to user *m* of iSC *l, i.e.* $a_{m(l),n} = 1$ if user *m(l)* is allocated sub-channel *n*. Hence, the optimization problem corresponds to a weighted sum-rate maximization problem in presence of inter-cell interference subject to power constraint of $P_{l,\max}$ per node *l* as in (4.40) and orthogonal allocation at intra-cell as in (4.41).

Approach

a. Proposed Graph-based Framework

The generic weighted rate maximization problem as described in (4.38) is a non-convex optimization problem with non-linear constraints and NP-hard. In this work we investigate a holistic graph-based solution that targets improving the cell spectral efficiency via better dynamic reuse across the cells in a networked small-cell environment. This involves a locally-centralized graph-based Inter-cell Interference Coordination (ICIC) via user partitioning across different clusters. Subsequently, the resource allocation policy is formulated as weighted sum rate maximization (WSRM) to optimize system performance in terms of both throughput and fairness.

Inter-cell interference is managed through an adaptive graph-based ICIC scheme, which combines graph-partitioning and local search concepts to provide near-optimal interference isolation between users of different cells. Subsequently, the adaptive clustering of users based on their mutual interference levels results into an SNR maximization problem where optimal resource allocation is accommodated by RANaaS for clusters of users aggregately.

- **Graph-Construction:** An interference graph *G (V, E)* is created, that consists of *V* vertices that correspond to the users in the system and *E* edges that show the downlink interference conditions between users. An edge between them logically shows the level of signal degradation to both users assuming they utilize the same resource part. This graph is a weighted un-directional graph that connects all the users in the system. This interference graph is constructed in the RANaaS. For the graph construction, we use a metric corresponding to the relative channel qualities for each pair of users. This metric encapsulates channel statistics to represent the worst case interference that each pair of users can experience at a specific sub-channel (path loss, shadowing effect and multipath fading).
- **Graph-partitioning:** Having formed the interference graph, we then focus on the graph-partitioning phase, proposing a novel formulation for the efficient partitioning of users into clusters. WSRM problem can be mapped into the problem of optimal partitioning of users into each cluster via employing the already created weighted interference graph. Such partitioning can be decomposed into a set of graph-based sub-problems, termed as Minimum Path Selection (MPS) per sub-channel. The MPS sub-problem is then defined as an integer programming problem and an exact solution is derived using Branch-and-Cut method. Due to high complexity of the problem, we propose an adaptive graph-

theoretic solution framework, wherein we provide near-optimal heuristic approaches. Subsequently for each sub-channel, from the derived candidate cluster set, we perform multi-cell resource allocation on a per cluster-basis, denoted as Multi-cluster Resource Allocation (MCRA) such that the weighted sum rate is maximized. Following, these two problems are further discussed:

Minimum Path Selection (MPS): Minimum Path of order v for sub-channel n, denoted as $\Phi_n^*(v), \forall n \in N, \forall v \in (0, L]$ is the minimum path of size v that traverses the interference graph where it includes only one vertex per disjoint set (iSC l). Mathematically, the problem of MPS, given an interference graph comprising L disjoint sets can be formulated as:

$$\Phi_n^*(v) := \min_B \sum_{i \in M_T} \sum_{j \neq i \in M_T} c_{i,j,n} b_{i,j,n}, \forall n \tag{4.42}$$

Subject to:

$$\sum_{j \neq i \in M_T} b_{j,i,n} \leq (v-1)\, y_i, \forall i \in M_T \tag{4.43}$$

$$\sum_{i \in M_v} y_i = 1, \forall v = 1, 2, .., L \tag{4.44}$$

$$b_{i,j,n} \in \{0,1\}, \forall i, j \neq i \in M_T \tag{4.45}$$

$$y_{i,n} \in \{0,1\}, \forall i \in M_T, \tag{4.46}$$

where $c_{i,j,n}$ is the edge cost between a pair of users for sub-channel n, $b_{i,j,n}$ is the binary variable corresponding to the allocation decision for the sub-channel n to both users i,j of different iSCs. Moreover, y_i is the set of auxiliary slicing variables which shows if a vertex is visited (is equal to 1) or not (is equal to 0). The constraint (4.43) requires the number of edges incident with a vertex to be either 1 (if i is visited) or 0 (otherwise) and constraint (4.44) ensures that exactly one vertex (or node) per disjoint set is visited. This problem will be solved using an exact method (branch-and cut) as well as a proposed heuristic approach.

Multi-Cluster Resource Allocation (MCRA): Having formed the L-1 minimum cost paths per sub-channel, at this stage, we try to identify the optimal MPS per sub-channel taking into account the already relaxed power constraint. This problem can be represented as finding the minimum path of size v for which the total costs of the users comprising the minimum path is minimized. This is equivalent to finding the optimal v for which the WSR of the users in the minimum path is maximized.

b. Proposed solution for MPS

In this section, we discuss the solution framework for the MPS sub-problem, which was defined in the previous section as an Integer programming problem. Firstly, we derive an exact solution to this problem using the branch-and-cut algorithm [33]. Furthermore, due to the high complexity of the enumerative solution we also propose a near-optimal heuristic algorithm to solve this problem efficiently for large graphs.

Branch-and-cut Exact Approach: The exact solution of the MPS problem follows a branch-and-bound scheme, where lower bounds are computed by solving a linear program relaxation of the problem. This relaxation is iteratively tightened by adding valid inequalities to the formulation according to the cutting plane approach. This method is known as a branch-and-cut algorithm and is thoroughly described in [33] for the case of the integer programming problem.

Proposed Heuristic Approach: Due to NP-hardness of MPS sub-problem, it is crucial to seek heuristic solutions to address the problem in an efficient manner. Therefore, we propose such a solution comprising three key steps:

- Selection of Representatives: This step enables the selection of one representative node corresponding to each cell. This representative node is the user with the best experienced signal quality towards its serving iSC.
- Generation of multiple minimum-cost paths for each representative: Thereafter, from each representative the minimum-cost paths are calculated. The minimum cost path is calculated by taking the intra-path sum weight, i.e. the sum of all the edges' weight combinations for the nodes

composing the path. Note here that the minimum cost paths that are generated can be sub-optimum solutions due to falling in local optima. In this stage, we generate a population of feasible solutions with path size *l*. The same procedure is repeated for all the representatives. As duplicate paths might be generated in this process, those are to be excluded from the feasible solution set at the end of this step.

- Selection of Minimum path: In this step, from the set of feasible solutions generated in the previous step, we select the minimum path of size *l* as the path with the lowest intra-path sum-weight among them.

c. Proposed Solution for MCRA

Considering *l* as a variable for MPS sub-problem, we can obtain *L-1* minimum paths in the aforementioned graph consisting of $V_1,V_2,...,V_L$ disjoint subsets. Therefore, the problem of the optimal partitioning of users into a cluster can be seen as finding the optimal *l* for which the weighted sum rate of the users comprising the minimum path is maximized. This problem is performed for all sub-channels independently (*N* times), resulting in *N* clusters of users in which the WSR gets maximized. One challenge here is that it is not possible to determine in advance the power level per resource, due to the fact that the proposed scheme may provide different resource utilization per cell. Therefore, the aforementioned challenge requires an iterative power allocation algorithm on the top of the graph-partitioning based channel assignment. We apply the optimum power allocation as derived in [34]. This algorithm is an iterative power allocation scheme dealing with the problem described above. The concept in this algorithm is to iteratively adjust the power-level per resource for each iSC based on the cluster channel assignments since the number of used resources per iSC is unknown in advance.

4.5.2 Implementation of CT in the iJOIN architecture

The proposed graph-based ICIC mechanism requires some signalling exchange between iSCs and RANaaS entity. Initially all iSCs within veNB receive channel state information per UE (CQIs). This information is then forwarded to RANaaS as part of RRC measurement reports. Based on the channel states of all UEs, RANaaS allocates PRBs such that the WSR per cell is optimized. This decision is then fed back to the corresponding iSCs and user data are transmitted to the allocated users respectively.

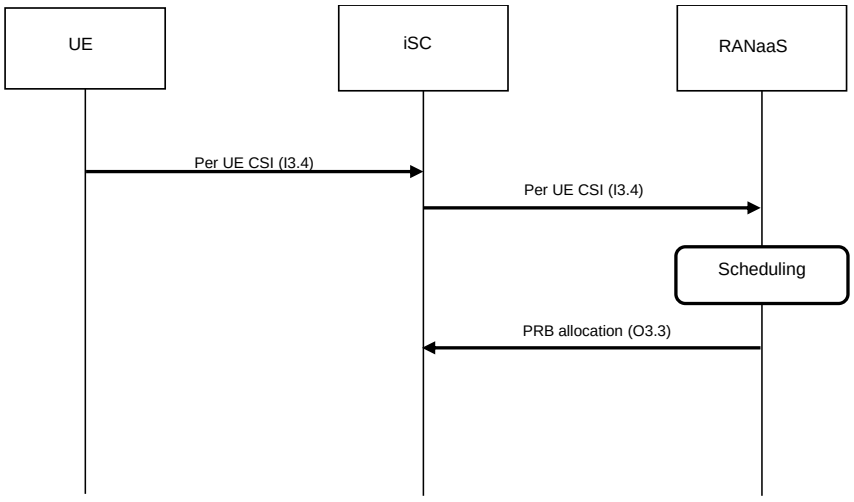


Figure 4-28 Message sequence chart for Cooperative RRM as proposed in CT3.5

4.5.3 Evaluation of the CT

Performance evaluation is done by means of numerical system-level simulations. The system consists of a dense iSC deployment and a local controller (i.e. RANaaS). The small-cell deployment used in this study is a 3x3 grid of apartments. In this deployment, each iSC and 4 users are randomly distributed in each apartment. The 5x5 path loss model (indoor networked small cells) is used to evaluate our model in a dense deployment of small cells derived from 3GPP [32], where we run Matlab Monte Carlo simulations using the simulation parameters of the 3GPP's specification for indoor networked small cells. The framework will also be evaluated using a common iJOIN simulation scenario as discussed in Section 5.

Compliance with iJOIN objectives

The outcome of this work shows promising results when tested in scenarios consisting of dense indoor small cells. The ICIC framework aims to enhance small cell's spectral efficiency / throughput by jointly scheduling users of different cells and at the same time to mitigate inter-cell interference by keeping the outage probability in low levels.

Description of the baseline used for the evaluation

The comparison of our proposal is performed using benchmarks from the state-of-the-art literature; and especially other competitive graph-based schemes and the full interference scenario. All these benchmarks use Proportional Fair scheduling and have been tested under the same conditions - weights to ensure fairness for the comparison.

Discussion of results of the CT

For evaluation purposes our proposal is compared with the case where interference management is only available via Intra-cell Scheduling (Proportional Fairness) in Reuse-1 and Reuse-3 scenarios. Furthermore, we compare our proposal towards competitive graph-based Dynamic ICIC approaches that were introduced in [35]-[36]. The following figures show the gains of our proposal taking the CDF of downlink SINR (Figure 4-29-left) and the CDF of cell spectral efficiency (Figure 4-29-right) as a performance metric for the achievable per cell spectral efficiency.

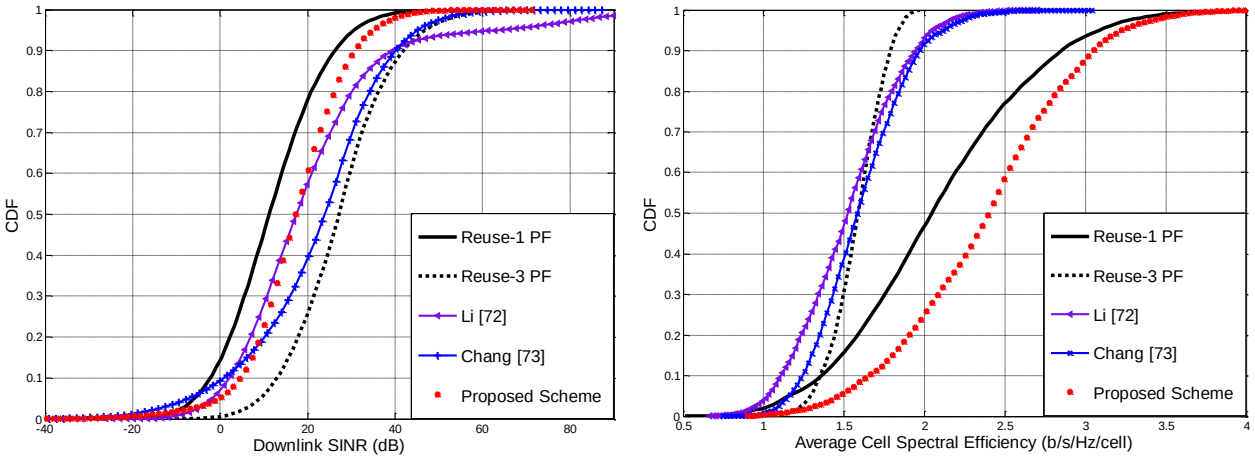


Figure 4-29 CDF of DL SINR (left) and CDF of cell spectral efficiency (right)

As can be seen in Figure 4-29-right, our proposal using discrete edge weights, shows significant gain over the benchmarks targeting the median and mean of the CDF of the average cell spectral efficiency. In particular, for both mean and median, we observe an improvement of 18% over Reuse-1 and more than 45% over the rest benchmarks.

Another interesting metric is the CDF of the downlink SINR. Here, 2.2dB (threshold for BPSK to achieve reasonable un-coded BER) is chosen as the threshold for the outage probability. By this, we can observe in Figure 4-29-left that our scheme has 6.5% outage probability and outperforms Reuse-1 PF (20% outage), [35] and [36] (~11% outage). On the other hand, it has higher outage than the Reuse-3 scheme (1%).

Furthermore, we aim to capture the impact of resource granularity (i.e. varying the number of sub-channels) as well as cell deployment on the performance of the proposed algorithm. Figure 4-30 illustrates the cell throughput for different resolutions of resource blocks (2 to 12) for two deployment strategies: regular small cell deployment where iSCs are located at the centre of each apartment, random small cell deployment where iSCs are randomly positioned per apartment. For each strategy, we demonstrate 2 curves: one related to the samples collected from only centre apartment (as the worst case scenario) and another related to averaging over samples collected from all apartments in the grid.

As illustrated in Figure 4-30, the random deployment shows lower performance compared with regular deployment for similar resource resolutions. The gap is larger for worst case scenarios collected from central

cells. Nevertheless, the curves are increasing function of resource resolution as it is evident by superior performance in the higher number of resources.

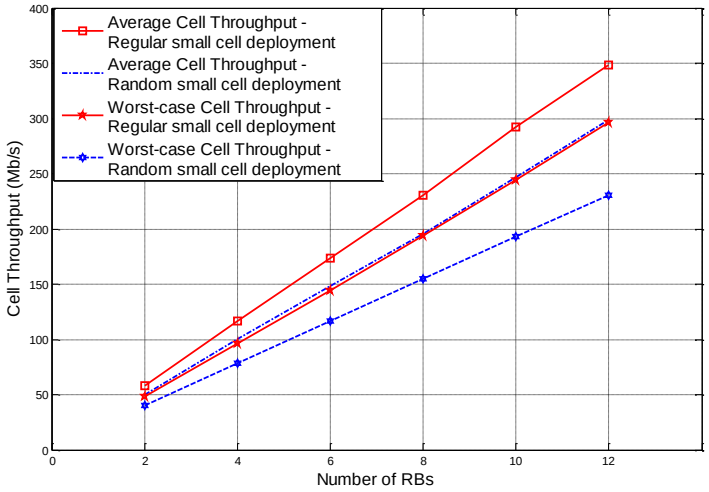


Figure 4-30 Cell Throughput Comparison for random vs. regular deployment

4.6 CT 3.6: Utilization and Energy Efficiency

This section introduces the key performance indicator Utilization Efficiency (UEff) which is of particular relevance for the iJOIN system. In addition, this new metric is linked to Energy Efficiency (EEff), which indeed is closely related to UEff. Initially, both metrics are defined and then approaches are introduced which improve the utilization efficiency.

4.6.1 Technical description

Measurements in operator networks reveal [50] that 20% of all base stations carry 50% of the overall traffic, meaning that the average utilization ratio is less than 40%. The main reason for this phenomenon is a wide deployment of macro-cells to achieve a high coverage, and the network dimensioning trimmed to peak traffic demands, meaning that a large fraction of deployed resources are underutilized. iJOIN aims at increasing this utilization by means of its two technology pillars, i.e., RANaaS decentralisation and joint RAN/backhaul design.

Utilization efficiency is defined as a metric expressing how well the utilized resources are employed to achieve a given performance metric. Therefore, high UEff means the following:

- The system (such as a network) is highly utilized, and therefore not over-provisioned.
- The system is capable to exploit utilized resources efficiently to provide the desired output, such as cell throughput or other targeted metrics.

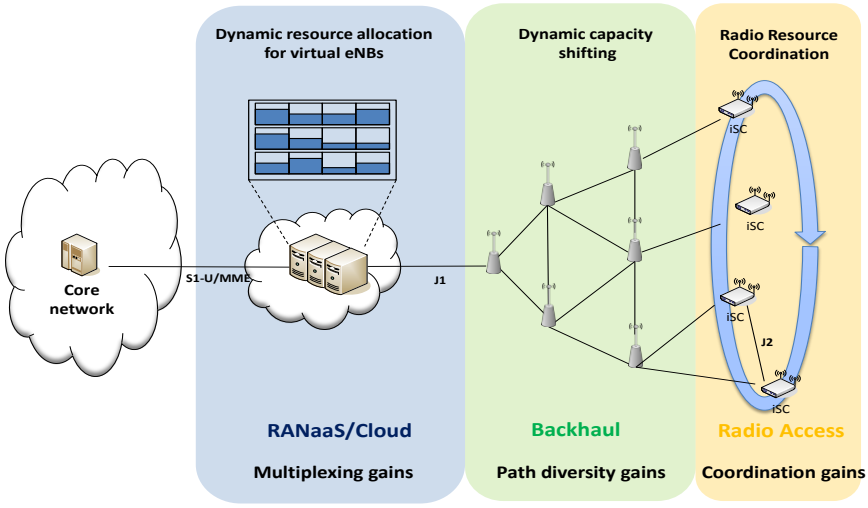


Figure 4-31: Utilization gains in different network domains

Figure 4-31 shows an example of how different resource allocation techniques in different iJOIN network domains can lead to different types of gains (e.g. multiplexing, diversity and coordination gains). It also illustrates a fundamental problem of defining a network-wide metric for utilization efficiency: different network domains (i.e. RANaaS, backhaul, radio access) utilize different types of resources (e.g. CPU cycles, link bandwidth, radio spectrum), such that a simple summation of domain-specific metrics is in general not possible. We define the total UEff of a system as following:

$$\eta_U = \frac{\sum_{d \in D} \alpha_d u_d}{|D|} \tag{4.47}$$

where α_d is a scaling factor s.t. $\sum \alpha_d = 1$, and u_d is the *domain utilization* for the considered domain, with D as the set of network domains (e.g. RANaaS, backhaul, RAN).

The definition of the domain utilization u_d depends on the resource of interest. As described in (4.47), different network domains have in many cases different resources. However on a more abstract level, resource normalization can be applied across network domains. We identified the following resource classes which will be investigated in more detail:

- Bandwidth/capacity resources. The domain utilization is defined as

$$u_d^B(X) = \frac{B_{mean,d}(X)}{B_{cap,d}(X)}, \tag{4.48}$$

  where $B_{mean,d}(X)$ is the average measured data rate and $B_{cap,d}(X)$ is the corresponding outage or theoretical maximum capacity of the system. The parameter X depends on the investigated network scenario and can be the number of cells, user arrival rate, etc.

- Computational resources. Here, the domain utilization is defined by

$$u_d^C(X) = \frac{C_{mean,d}(X)}{C_{outaged}(X)}, \tag{4.49}$$

  where $u_d^C(X)$ is the ratio of expected computational demand and provided computational resources, depending on the number of cells in the scenario, X. The latter is the outage complexity which is defined as the amount of computational resources to make sure that a per-cell computational outage ε is not exceeded. Both are defined through an analytical framework which has been partially described in [63]. This framework resembles the characteristics of computational load of a 3GPP LTE uplink decoder.

Energy efficiency is defined as a qualitative metric expressing the impact on power (or energy) consumption of the transition from the standard 3GPP eNB architecture to the iJOIN model. In general, the full scenario needs to be evaluated to make fair comparisons; in other words, the metric should be evaluated taking concurrently into account all the applied candidate technologies and/or functional split instances at once, since the effects of their simultaneous application is not necessarily additive.

Hence, given a specified physical configuration and scenario (fixing scenario parameters like spatial coverage, number of served UEs, delivered area throughput, etc.), on one hand we have the standard 3GPP LTE configuration, where the power consumption is essentially given by the eNB's component plus the backhaul component:

$$\sum_{n=1}^{N_{eNB}} P_{eNB\_n} + P_{BH} \tag{4.50}$$

On the other hand, in case of a iJOIn, veNB-based configuration, there are more single items building up the total power consumption:

- The energy spent in the iJOIN small cells for the computational part (excluding radio functions), given as sum of the DSP energy spent in each active iSC:

$$\sum_{n=1}^{N_{iSC}} P_{iSC\_n} \tag{4.51}$$

- The "useful" energy spent inside the cloud, meant as power consumption of the servers in the datacentre spent to run the computational workload related to the RANaaS decentralized processing functions:

$$\sum_{n=1}^{N_{Server}} (P_{RANaaS\_n}) \tag{4.52}$$

- The energy spent as "due overhead" inside the cloud datacentre, i.e., cooling and conditioning, Uninterruptible Power Supply (UPS) and other facility related consuming equipment (lights, etc), described with good approximation applying the datacentre parameter PUE [66] (Power Utilization Effectiveness)[4]:

$$(PUE - 1) * \sum_{n=1}^{N_{Server}} (P_{RANaaS\_n}) \tag{4.53}$$

- The energy spent in the backhaul links (logically separated by the iSCs for this specific metric):

$$P_{\mathrm{Bh}} = \sum_{n=1}^{N_{\mathrm{iSC}}} P^{n}_{\mathrm{switch}}\left(y_n\right) + N^{n}_{\mathrm{mw}} P^{n}_{\mathrm{link}}\left(y_n\right)$$

, where

- P^{n}_{switch} is the power consumption of switches at each iSC aggregating traffic from other iSCs in case more than one backhaul link originates at the reference iSC;
- N^{n}_{mw} is the number of microwave antennas at each iSC n;
- P^{n}_{link} is the power for transmitting and receiving the aggregate backhaul traffic at each iSC;
- y_n is the load at each iSC n.

(4.54)

Hence, the comparison is between (4.50) versus the sum of (4.51) to (4.54). The most significant weight on the metric assessment is expected to come from the marginal consumption caused in the RANaaS point(s) of presence compared to the processing induced consumption in the standard eNB cells. Actually, even the backhaul consumption can't be taken equal by default in the two configurations: the presence of a J1

interface, and the backhaul adaptation features present in the iJOIN architecture, make the traffic crossing the backhaul change, consequently the power consumption will not be the same. Notwithstanding this, the major difference is reasonably expectable to come from the computational functions.

4.6.2 Evaluation of the CT

Computational Effort of RAN Functions

For RANaaS, the computational demand of RAN function execution is of special interest since the question of feasibility of the functional shift towards the centralized network entity needs to be answered for different functional split configurations. One of the main impact factors on the computational demand is the modulation and coding scheme (MCS) selection, which determines together with the SINR the number of transport block decoding iterations on the receiver side. Figure 4-32 shows an example of the computational effort for vs. the instantaneous SINR under the assumption of block fading. The "spiky" behaviour stems from MCS switching, and indicates that significant diversity gains in case of centralization can be expected. The analysis is based on the formulation of a complexity model for forward error correction which is described in detail in [63]. The shown curve was generated under assumption of a target BLER of 10% on the first HARQ transmission.

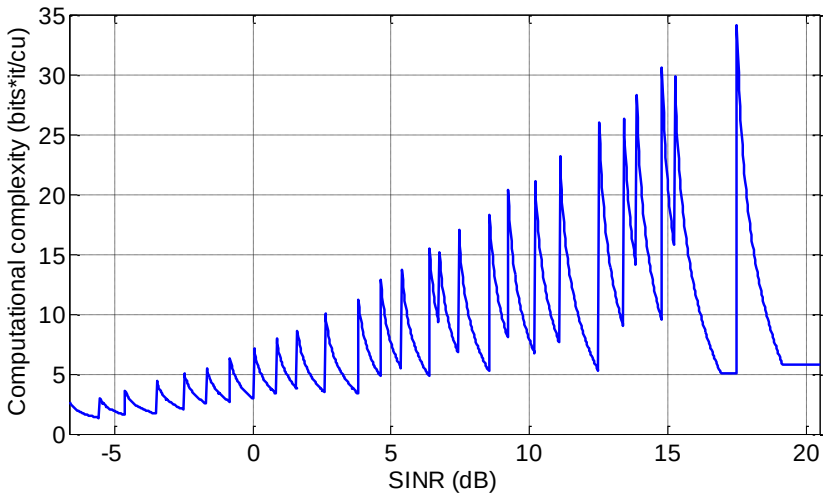
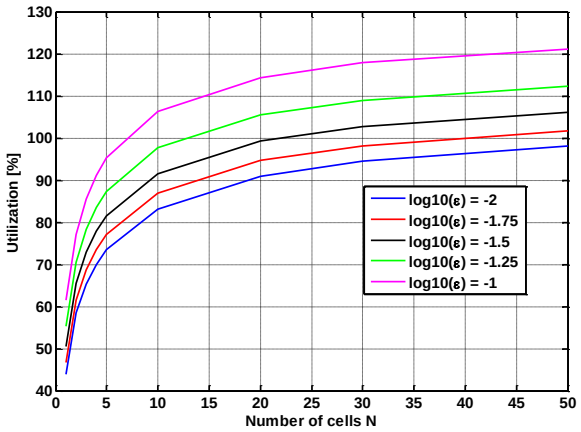


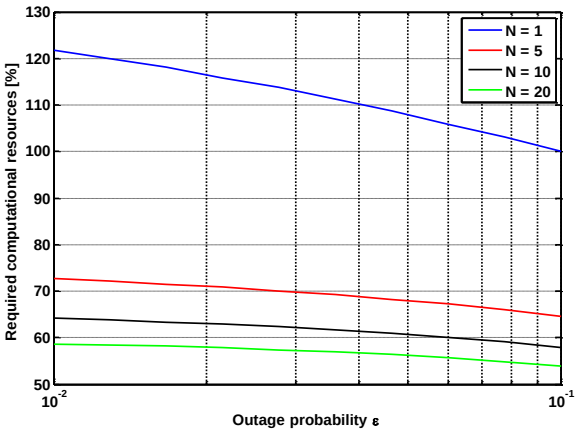
Figure 4-32: Computational effort as a function of the SINR

Analytic evaluation of computation utilization efficiency:

Based on this framework, the expected utilization of a centralized processor for different number of cells and depending on the outage is shown in Figure 4-33. For these results typical LTE parameters including actual SNR link-adaptation thresholds have been used. Furthermore, a Rayleigh fading process is assumed with an average SNR of 10dB.



(a) Utilization as a function of number of cells



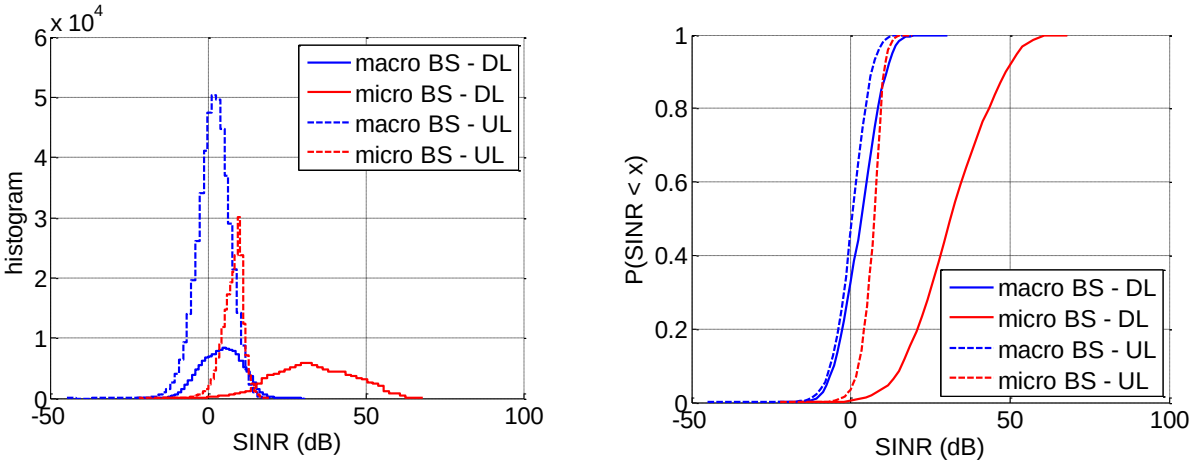
(b) Required outage complexity as a function of the per-cell outage probability

Figure 4-33: Computational utilization efficiency

From Figure 4-33(a) we can see that for a large number of centralized base stations, an expected utilization of more than 100% is achieved. This implies that less computational resources than the expected overall computational demand are provided. This is due to the fact that the system is optimized such that a per-cell outage probability is not exceeded. We can observe that this effect depends strongly on the chosen outage probability, e.g. for a computational outage of 10% already 7 centralized base stations would exceed the provided resources while for a computational outage of 1% more than 50 base stations need to be centralized. This utilization performance curve will be helpful to dimension the centralized resources accordingly and to design the resource scheduler. Based on the actual communication resource demand (throughput) also the computational resource demand (processing) can be scheduled, and vice versa.

System-level evaluation

For further evaluation of the computational aspect of UEff a calibrated system-level simulator compliant with 3GPP requirements is used. Channel fading traces are obtained in a 3-tier, wrap-around hexagonal 3-sectored layout with the IMT-Advanced spatial channel model [64], heterogeneous network deployment with clustered small cell and mobility/hand-over modelling. The computational complexity demand is calculated with a link-level implementation of the LTE turbo decoder and rate matching algorithm for an error rate of 0.1 for the first transport block transmission in the HARQ protocol. The evaluation scenario corresponds to the "square" common scenario defined in [15] depending on the parameterization.



a) histogram of SINR for macro and micro, UL and DL signals

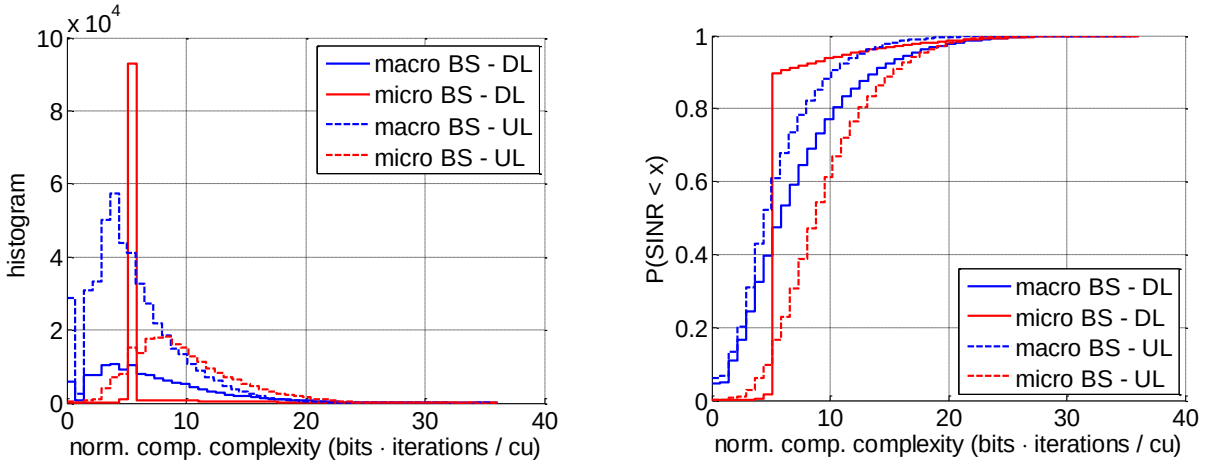
b) Cumulate distribution function of SINR

Figure 4-34: SINR distribution in a HetNet scenario

Figure 4-34 shows the SINR distribution in this scenario. It can be observed that UEs attached to iSCs (micro BS) experience a significant better SINR especially in downlink. The uplink has in both the macro and in the micro case a smaller dispersion compared to the downlink, which means that MCS changes can be expected to occur less frequently from a system-wide perspective. Note that open loop uplink power control as specified in [65] is and Round-Robin scheduling is implemented.

The corresponding histogram and CDF of normalized computational complexities are shown in Figure 4-35. Due to the high SINR advantage of UEs attached to iSCs in downlink, they have a high chance that the highest MCS is selected and decoding is correct without retransmission and deep iteration by the turbo decoder. The strong peak at computational complexity value of 6 in Figure 4-35(a) is caused by this phenomenon, corresponding to an SINR of 18dB or higher in Figure 4-32.

The main direction of interest for computational complexity is the uplink, since otherwise decoding takes place in the UEs which are not subject to centralization for obvious reasons. Here it can be observed that micro BSs have a higher demand for computational complexity than macrocells. The reason is that on average, the computational complexity tends to increase with the SINR until a certain maximum value, as also shown in Figure 4-32.



a) histogram of normalized computational complexities

b) Cumulate distribution function of normalized computation complexity

Figure 4-35: Distribution of per-subframe normalized computational complexities

From the perspective of UEff the characteristics of the computational complexity (CC) function is not beneficial, as it indicates that to avoid computational outages, significant over-provisioning of resources is necessary. It is therefore interesting to investigate the impact of resource management, centralization and the corresponding multiplexing gains, if any.

Figure 4-36 shows a trace of the total cell computational complexity, and the corresponding number of UEs in the cell. It can be observed that a dependency exists (trivially in case if there are no UEs in the cell, but also e.g. at time index 3.73*10^4, where three UEs arrive at a cell). However, a strong correlation between the number of UEs and complexity cannot be necessarily concluded. The reason is that the main impact factor is the SINR which leads to potentially very strong changes of the CC function with small changes by its value.

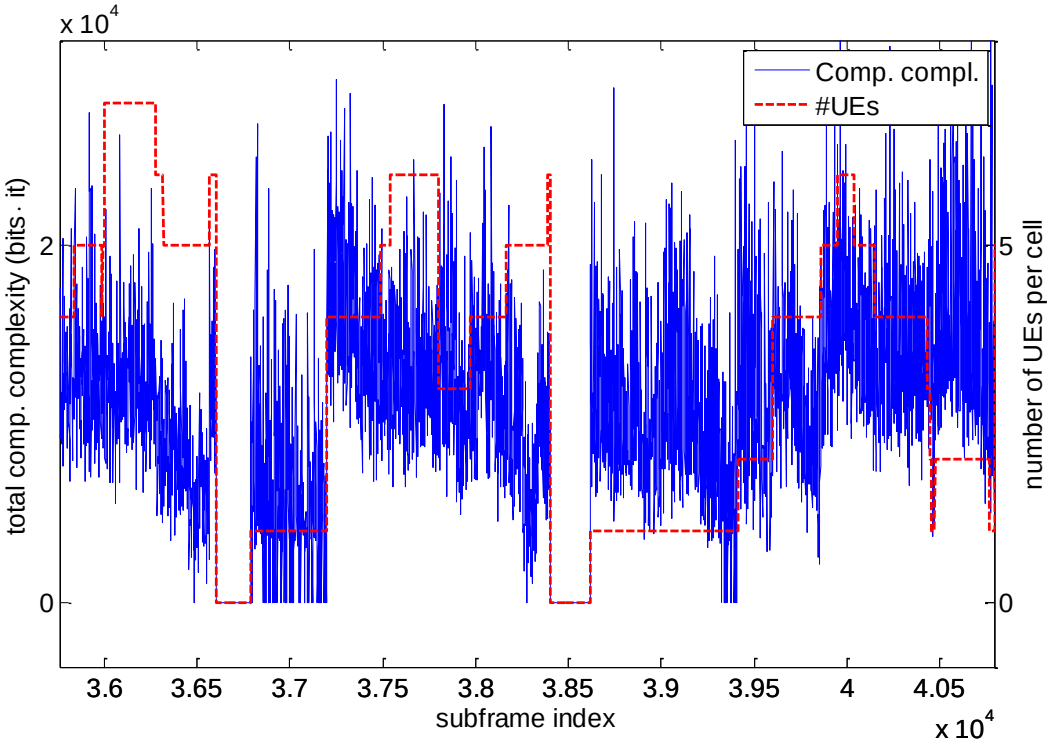


Figure 4-36: Trace of total computational complexity and number of attached UEs per cell (macro cells)

This is further illustrated in Figure 4-37, which shows the mean per-cell CC depending on the number of transmitting UEs. A correlation between the mean CC and the UE/cell density is not observable; the delta between different communication directions (UL/DL) for the same cell type stems from UL and DL having different mean value of SINR.

This result implies that from cell view, multiplexing gains may origin from a reduction of the dispersion of the total CC per-slot. To verify this, further simulation campaigns will be conducted such that statistically relevant results can be generated for the CC distribution and standard deviation depending on the number of UEs.

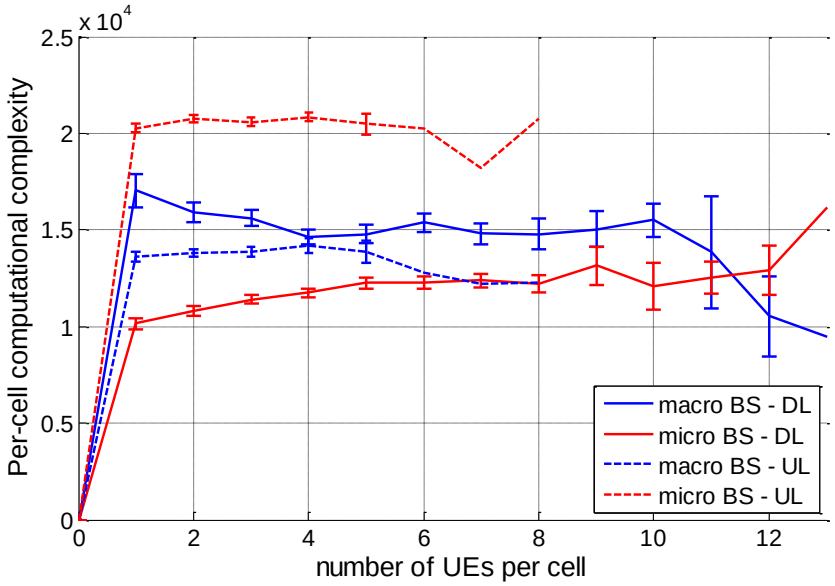


Figure 4-37: Per-cell computational complexity vs. number of attached UEs

4.7 CT 3.7: Radio Resource Management for Scalable Multi-Point Turbo Detection

4.7.1 Technical description

Scenario

In a dense small cell deployment, one user equipment (UE) can more easily see other small cells in addition to its serving one (especially if it is at the edge of the cell). Under such deployment, if co-channel deployment is used due to limited spectrum, classical approaches to improve the transmission quality tend to create orthogonality in the frequency domain for OFDM-based systems, e.g., through soft or fractional frequency reuse patterns among neighbouring small cells. By nature, these frequency partitioning schemes reduce the spectrum available for transmission, meaning less maximum throughput achievable in theory.

By scheduling the (edge) users on the same resources and exploiting the created interference as a source of information in each concerned small cell, it should be possible to improve the (aggregated) uplink throughput of the system, as “more” spectrum and diversity are made available to the users.

To deal with the created interference among the “aligned” users in the uplink, we rely on a scalable form of the turbo detection principle. Indeed, the turbo detection allows significant performance improvement [14] by relying on the information exchange (extrinsic log-likelihood ratios) between the detection stage and the decoding stage in an iterative way. However, one drawback of such iterative processing is the computational cost which increases linearly with the number of streams per users and the number of users involved in the detection.

In this scalable form, the turbo detection processing is either performed centrally at the RANaaS data centre (multi-point turbo detection - MPTD) or locally within each involved iSC (single-point turbo detection – SPTD). A radio resource management algorithm is needed to determine which iSCs and which UEs will benefit from this advanced multi-user detection. In both SPTD/MPTD cases, this centralised RRM algorithm will always be running in the RANaaS data centre on an iSC demand basis. It will provide the set of iSCs and UEs to be involved in the turbo processing process, thus giving a “long-term” scheduling framework for each involved iSC. The “short-term” scheduling will take place normally at the iSC level but under this framework.

If link-level simulations in particularly severe conditions clearly showed an advantage of both form of turbo detection [14] [43], the benefit in a large scale deployment has yet to be determined at the system level. To do so, we will investigate a deployment scenario targeting the airport/shopping mall common scenario [15] conditions with indoor dense hotspots operating on the same channel.

Figure 4-38 shows the small cell deployment scenario envisaged. Solid lines represent the minimum requirements assumed on the interface (High Quality, Medium Quality, Low Quality related to the bandwidth/latency capability of the link). Since part of the scheduling will be done in the RANaaS platform, the J1 interface should be sufficient to support such operation to take place. Due to close deployment location, each iSC-iSC link is assumed to be of high quality in terms of latency. Such link will carry J2 or extended X2 message information.

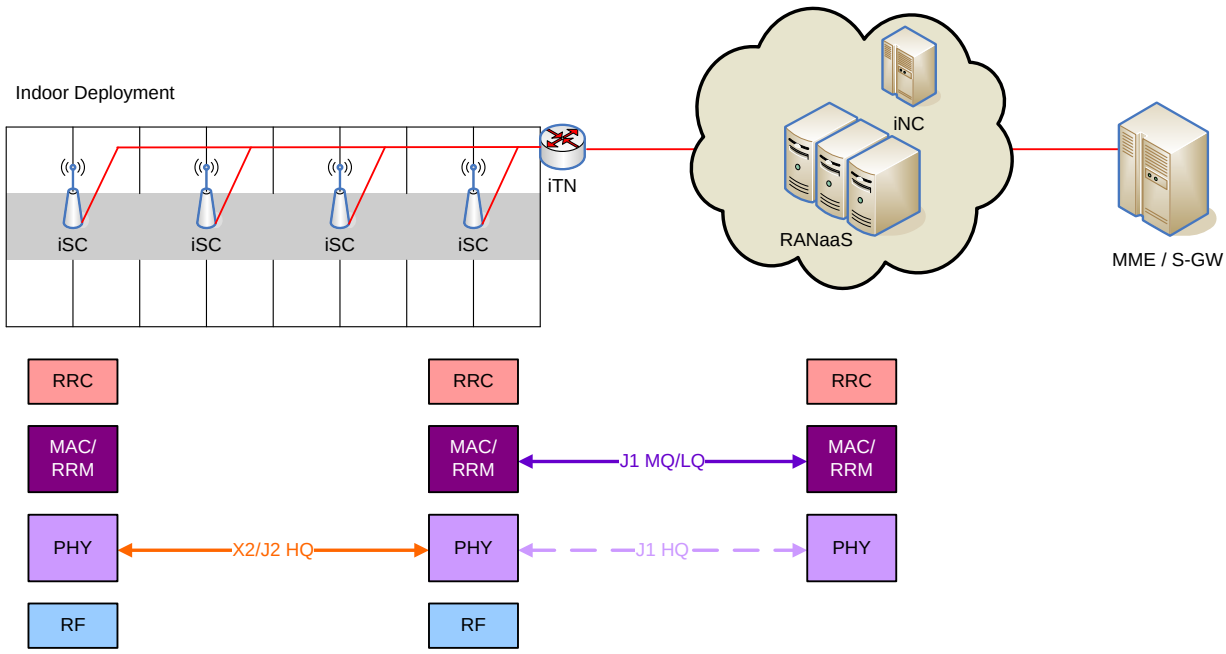


Figure 4-38: Scalable multi-point turbo detection scenario (solid lines are minimal requirements)

System Model

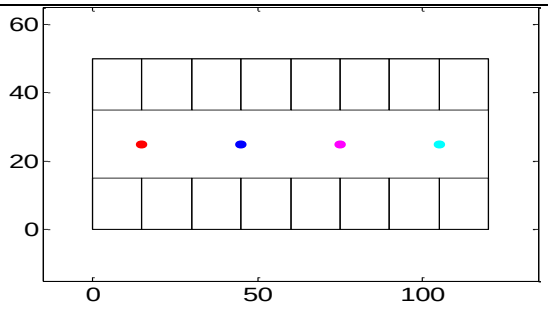
To assess the performance of the MPTD/SPTD approach, system-level simulations will be performed in the uplink direction. The setup and methodology is quite similar to the one defined for Scenario 3 (dense) in the 3GPP study item on the small-cell enhancements [21], using the extended Indoor/Hotspot (InH) layout from the ITU-R [22].

Through a Monte-Carlo approach, UEs will be uniformly dropped in the building and attached to an iSC based on the best received power criteria. Those UEs have full-buffer traffic to send in the uplink, with the same QoS priority. To simulate the load balancing operation, the same number of UEs will be attached to each iSC during their drop.

The uplink throughput per UE will be the main key performance indicator (KPI) monitored. Table 4-2 shows the main (static) assumptions used for system-level simulations, valid for both baseline and MPTD/SPTD investigations.

Table 4-2: System Level Simulation Static Parameters

LTE Parameters	
Bandwidth	10MHz
Frequency	$f_c = 2.6$GHz
Layout Parameters	



Block	Number	2 Rows of 8 Blocks
	Size	15m x 15m
Hall Size		20m x 120m
Deployment Parameters		
Number of iSCs		4 (fixed position)
Number of UEs		32 (random drop)
iSC Parameters		
Antenna	Number	2 (Uniform Linear Array)
	Spacing	0.5 λ (Wavelength associated to f_c)
	Polarization	Vertical
	Pattern	Omnidirectional
	Gain	0dBi
	Height	6m
Transmit Power		24dBm
Noise Figure		5dB
UE Parameters		
Antenna	Number	1
	Polarization	Vertical
	Pattern	Omnidirectional
	Gain	0dBi
	Height	1.5m
Transmit Power	Maximum	23dBm
	Minimum	-40dBm
Noise Figure		7dB
Propagation Parameters		
Thermal Noise Density		N0 = -174dBm/Hz
Channel Model		ITU-R InH [22]

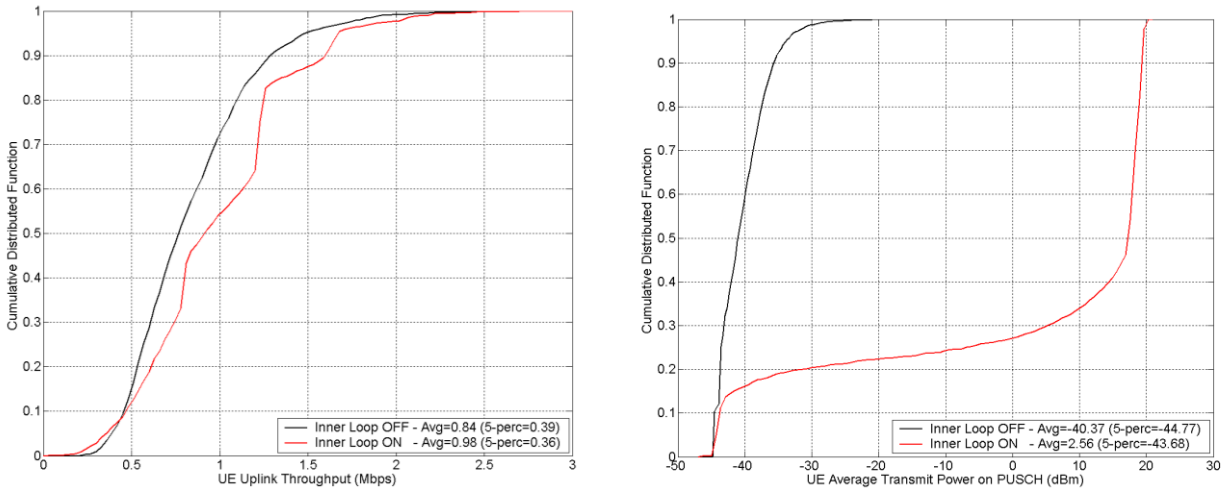
Line of Sight Probability (d is the iSC-UE 2D-distance in meters)		$P_{LoS} = \begin{cases} 1, & d \le 18 \\ e^{-\frac{d-18}{27}} & 18 < d < 37 \\ 0.5 & 37 \le d \end{cases}$
Pathloss	LoS	$PL = 16.9\log_{10}(d) + 32.8 + 20\log_{10}(f_c)$
	NLos	$PL = 43.3\log_{10}(d) + 11.5 + 20\log_{10}(f_c)$
Shadowing Std Dev	LoS	$\sigma = 3\text{dB}$
	NLos	$\sigma = 4\text{dB}$

Dynamic system level simulations with the baseline scenario are done with the main parameters given in Table 4-3.

Table 4-3: System Level Simulation Dynamic Parameters

Dynamic Parameters	
Number of RBs for PUCCH	2 (1 + 1)
Number of RBs for PUSCH	48
HARQ	8 synchronous HARQ processes. Chase combining
Overhead	DMRS assumed (12 SC-FDMA symbols available per frame) SRS not simulated
Power Control (outer loop)	P0 = -106dBm, α = 1.0 [23]
Fast Fading	ITU-R InH [22]
Traffic	Full Buffer
Scheduler	Equal resource repartition / CT3.7 scheduler
Physical layer abstraction	MIESM compression LTE-compliant AWGN look-up tables per MCS and RB

Only the outer-loop power control algorithm is used [23]. Due to the dense deployment of UEs, preliminary simulations have shown that applying an inner-loop power control every TTI increases a bit the performance but makes the UEs transmit with really high power as shown in Figure 4-39 where more than 45% of the UEs is almost transmitting at the maximum transmit power on the PUSCH channel. Cleverer inner-loop power control activation is for further study.



(a) **(b)**

Figure 4-39: CDF of the average UE UL throughput (a) and transmit power (b) with (red) and without (black) inner loop power control

Equal resource allocation is assumed, meaning that each UE will be scheduled on 6RBs since 8 active UEs are attached per iSC.

A compliant LTE system-level tool is used. Therefore, UEs scheduled at subframe *n* through the PDCCH will be allowed to transmit on the PUSCH at subframe *n*+4. They will receive at subframe *n*+8 a positive or negative acknowledgement through the PHICH as well as the next transmission grant. In case of a new transmission or a retransmission, this will occur on subframe *n*+12, leading to a synchronous use of each HARQ processes: with a 8ms round time trip (see Table 3-3), 8 HARQ processes are used to avoid any gap in the UE's transmission.

Approach

For the turbo detection to be possible (either locally or centrally), some users served by different iSCs need to be "paired", i.e., aligned on the same resources. The main idea behind the turbo principle is to use the interference coming from a different user as a source of information to decode the user of interest signal in an iterative process: the user signal is detected and its contribution in the receive signal is estimated and subtracted to improve the detection of another.

Ideally, there should not be a too strong difference in the contribution coming from each UE in the received power to avoid any saturation due to quantization. Such processing seems to be beneficial for the edge users. Intuitively, one can think that at a cell edge, one user sending data toward its serving cell without beamforming will also be detected by at least one neighbouring cell with a received power "comparable" to the one at the serving cell.

In our CT, the turbo-detection is triggered on an on-demand basis by the iSC and will only involve two UEs and two iSCs per request. The RANaaS will run the RRM algorithm in charge of pairing the users, i.e., deciding which UE/iSC will be paired with as well as the resource to be used. A "static" approach will be described here, the dynamic one (including the on-demand aspect) being for further studies.

In a "coordinated" manner, the iSCs report to the centralised RRM instance running in the RANaaS data centre, the list of users which see an iSC other than their serving cell within a given threshold. The measurement is based on the downlink received signal: the Reference Signal Received power (RSRP). To facilitate this measurement, we assume that each iSC is one cell and that neighbour iSCs have different Physical Cell Identities (PCIs) modulo 6. While this is a downlink measurement, it is reasonable to use it as an approximation to detect users which uplink transmissions may be received with "comparable" powers by the corresponding iSCs.

More precisely, let assume the following notations:

- S is the number of iSCs close to each other ($S = 4$ in our example);

- U_s be the set of users attached to the iSC s ($U_s = 8$ in our example);
- P_s^u be the RSRP value of the iSC s measured by the user u in dB;
- $\Delta_{threshold}$ is the RSRP difference threshold for the pairing.

Each iSC will send to the centralised RRM instance a subset of candidate users $U_s^{candidate}$ constructed such that:

$$U_s^{candidate} = \left\{ u \in U_s \,\middle|\, \exists s' \neq s, P_s^u - P_{s'}^u \leq \Delta_{threshold} \right\}. \quad (4.55)$$

Practically, each entry will have a user candidate, its dominant neighbour cell/iSC and the associated RSRP measure.

The centralised RRM algorithm will perform the following operations (high level description)

- **For each** iSC s
  - Sort all user candidate subsets based on the RSRP values
  - Set a list of available resource blocks representing the PUSCH space and mark them all free
- **end for each**
- **For each** iSC s and **until** $U_s^{candidate}$ is not empty ,
  - Extract/remove the first user candidate u_s from $U_s^{candidate}$ and identify its dominant neighbouring cell s'
  - **If** a user $u_{s'}$ exists in $U_{s'}^{candidate}$ which has s as dominant neighbouring cell, **then**
    - Extract/remove the user candidate $u_{s'}$ from $U_{s'}^{candidate}$
    - Form the pair $(u_s, u_{s'})$ as a candidate pair for turbo detection
    - Find consecutive resource blocks which are free for both iSC s and iSC s'
    - Mark those resource blocks as used by s and s'
- **end for each and until**

At the end of the processing, the centralised algorithm has decided which UEs should be paired and on which resources. Of course, this pairing does not take into account the quality experienced by the users on these resource blocks and is therefore, suboptimal. However, it does not require to feedback channel state information which could be rapidly bandwidth hungry and are really latency sensitive. This method only provides a large scale framework and the MAC algorithm will compensate the channel imperfection on these resource blocks by adapting the modulation and coding scheme to the actual conditions and optionally use the inner-loop power control if necessary.

4.7.2 Implementation of CT in the iJOIN architecture

No matter the functional split, the previous algorithm will always run in the RANaaS data centre. Based on the backhaul conditions, the physical processing associated to these users could be done either at the RANaaS (this will also include the lower MAC part, i.e., the short term scheduling) or locally in each iSC.

Functional Split A) Multi-Point Turbo Detection

In MPTD, the physical processing associated to the users is also done in the RANaaS. Figure 4-40 shows such functional split for a simple example. The box running in the RANaaS is in charge of the centralised RRM algorithm previously described.

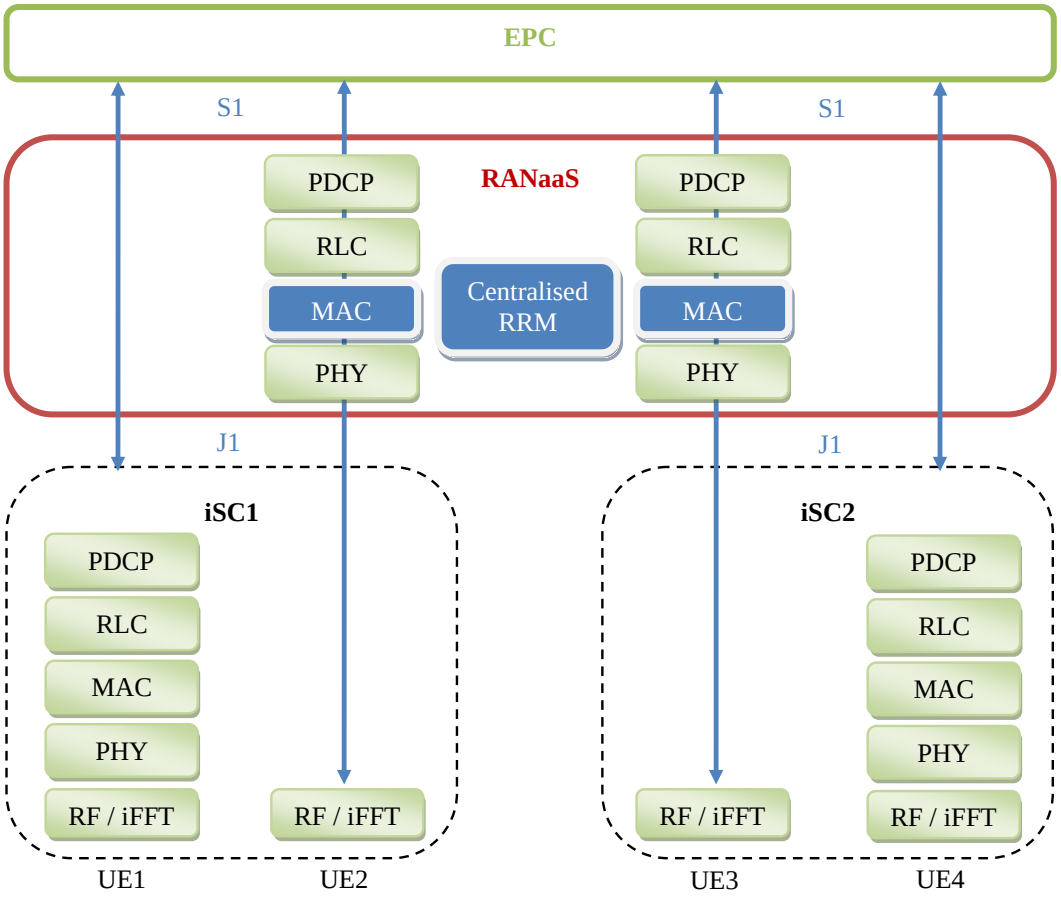


Figure 4-40: MPTD functional split when involving UE2/iSC1 and UE3/iSC2

In terms of integration in the iJOIN architecture, the Figure 4-41 shows the message sequence chart (MSC) supporting our centralised RRM algorithm when MPTD is possible. Based on measurements of their attached UEs (1a/1b), the iSCs can request the RANaaS to provide centralised RRM help by sending a list of candidates which RSRP satisfy the (configurable) threshold parameter $\Delta_{threshold}$ (2a/b). Based on these measurements, the centralised RRM tries to identify the possible pair of users (3) to involve in the turbo detection process. If a pair can be found (e.g., UE1/iSC1 and UE2/iSC2 in our example), the RANaaS gets the backhaul status from the two involved iSCs from the iNC (4 & 5) and decides whether MPTD or SPTD should be used, i.e., central processing at or distributed processing (local turbo detection process).

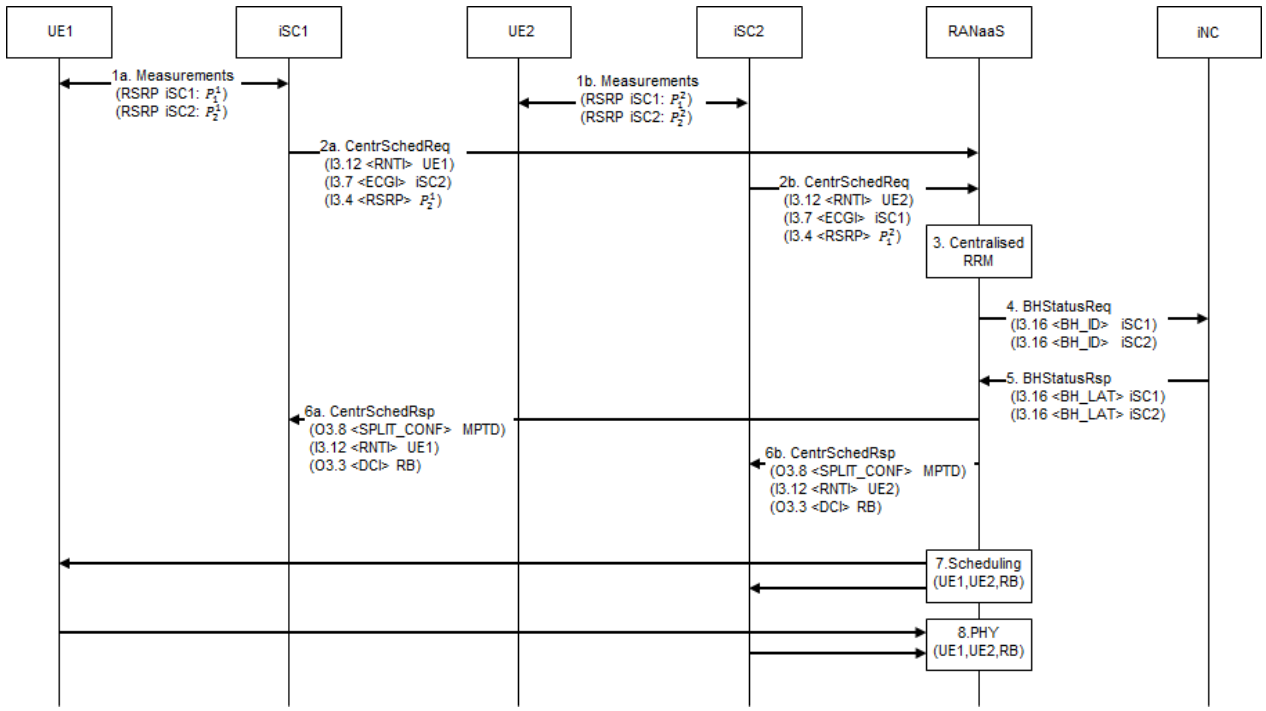


Figure 4-41: Message sequence chart for centralised RRM in case of MPTD

In the case where MPTD is possible, the RANaaS informs each iSC that it will perform MPTD for a given user, on a given list of resource blocks (6a/b). From now on, RANaaS will schedule the two users, i.e., derive the MCS to use on the configures resource blocks and send to the iSCs these uplink scheduling information to transmit on the PDCCH (7), while the iSCs will normally deal with their other users, exploiting the resource blocks not signalled by the RANaaS. When receiving transmission on the PUSCH and if the RANaaS has sent 4 subframe earlier an MPTD scheduling order on the PDCCH, the iSC will forward any I/Q samples signalled by the RANaaS as being used for MPTD (8) and perform a classical detection of the other resources blocks.

Functional Split B) Single-Point Turbo Detection

In SPTD, the physical processing is performed at iSC. To help the turbo detection, synchronisation is used between the two iSCs to signal to each other when a candidate user has a transmission scheduled. To do so the J2 must have a very low latency or a scheduling pattern may also be used (using a predefined set of subframes where SPTD should be used). Figure 4-42 shows the equivalent functional split.

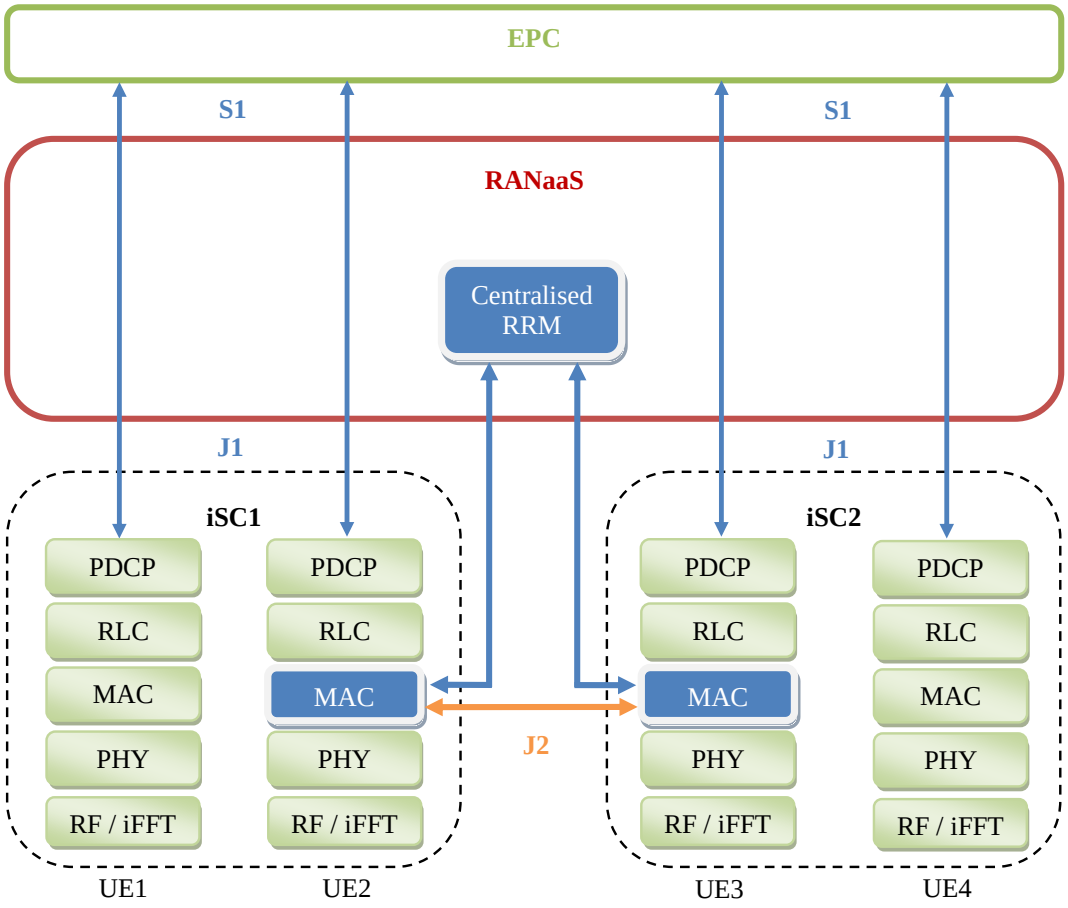


Figure 4-42: SPTD functional split when processing involving UE2/iSC1 and UE3/iSC2

In terms of integration in the iJOIN architecture, the Figure 4-43 shows the MSC supporting our centralised RRM algorithm when SPTD is possible. Up to the step (5), SPTD and MPTD share the same procedure.

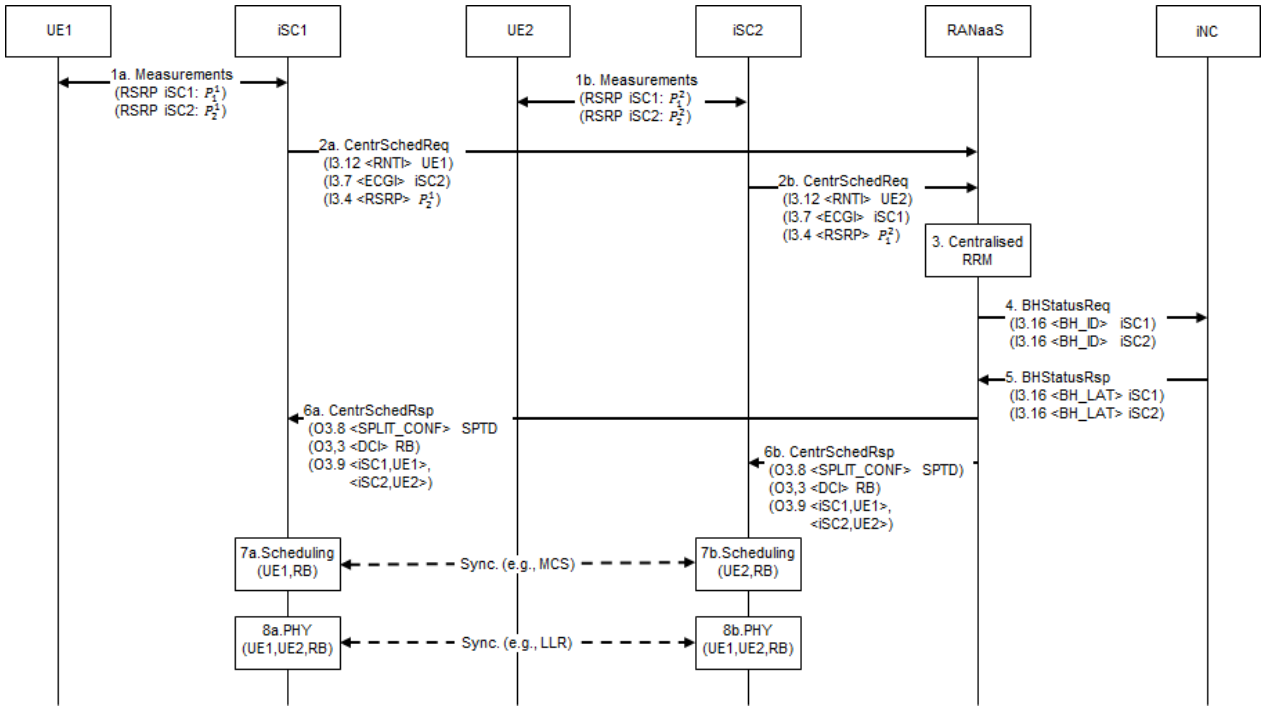


Figure 4-43: Message sequence chart for centralised RRM in case of SPTD

In the case where only SPTD is possible, the RANaaS informs the iSCs that they will have to perform SPTD for their given user on a given list of resource blocks (6a/b). From now on, each iSC will schedule its SPTD

users on the resource blocks signalled by the RANaaS (7a/b). Cooperation is needed between iSCs either through the J2 interface (optional synchronisation) or through the use of a scheduling pattern defined by the RANaaS in order for the iSC to know when (and with which MCS) the paired users will be. When a scheduled SPTD is received, the iSC detects both users through the turbo detection (8a/b) and will only deal with its attached users. Optional cooperation through the J2 interface may be envisaged for the physical processing, which is left for further study.

4.7.3 Evaluation of the CT

Compliance with iJOIN objectives

By scheduling users on the same resources and exploiting the created interference as an additional source of information, the area throughput should be increased in theory, addressing the first objective defined in iJOIN [15].

Usually Backhaul capacity is usually over-dimensioned compared to the RAN capacity, therefore, increasing the area throughput should also increase the utilisation efficiency of both the air and backhaul interface, addressing the fourth objective defined in iJOIN [15].

Description of the baseline used for the evaluation

The baseline scenario consists of a dense deployment of LTE Release 10 small cells and UEs in an indoor environment. The support of the X2 interface with associated standardised message [16]-[20] is not supported.

An equal resource allocation scheduling algorithm is used as a baseline, which schedule all UEs at each TTI. Full-buffer traffic is assumed, simulating a dense deployment and a high network demand. This scheduling algorithm allocates the same amount of RBs to all UEs, by choosing them randomly and contiguously among the available RBs in the PUSCH domain. The MCS is updated from the last received transmission SINR with a FER target of 10^{-1}.

A Minimum Mean Square Error (MMSE) receiver strategy with Interference Rejection Combining (IRC) is assumed at each iSC as a baseline.

Discussion of results of the CT

To represent the turbo detection receiver strategy, a perfect user cancellation is assumed when computing the SINR, i.e., when computing the SINR of user 1, the contribution of user 2 is not taken into account and vice versa. Such approach is really optimistic, but the turbo-detection tends to usually reach this bound (see [43]).

We compare the SPTD and the MPTD approaches to the baseline scenario. Backhaul constraints are not considered in these results, giving us the best case scenario. It has to be noted that for MPTD, the backhaul capacity of the J1 link is not really an issue as we do not forward I/Q signals for all subcarriers, but only for those where MPTD users are scheduled (subset of resource blocks). Also the signalling useful to derive the centralised RRM framework does not rely on channel state information but on long term measurements (RSRP), which needs far less capacity on the J1 link. However, the round-time trip on the J1 is assumed to be below 3ms to stay LTE-compliant (for the HARQ process).

If such time constraint cannot be satisfied, then SPTD should be used instead. No cooperation between iSCs is assumed during the physical processing. However, we assume that cooperation occurred during the scheduling stage. In particular, the iSC knows when paired user(s) are scheduled and with which MCS. This needs either cooperation through the J2 link or a pattern of transmission opportunities derived by the centralised RRM algorithm (encompassing for instance subframes when SPTD should be applied, MCS range to limit blind decoding, etc). Such approach is left for further studies.

We choose a threshold value of $\Delta_{threshold} = 6\text{dB}$ for the user candidate set construction. With 32 active users, we have on average 8.87 users being paired with each other by the centralised RRM algorithm. A lower $\Delta_{threshold}$ value will lead to fewer candidates while a higher value will increase the candidate set size.

Figure 4-44 compares the cumulative density function (CDF) of the small cell uplink throughput for the baseline, SPTD and MPTD. We note that a centralised solution (MPTD) offers better performance than a local one (SPTD) without cooperation, while both options outperform the baseline approach. In particular the

average throughput equals to 20.9Mbps for the MPTD, 19.5Mbps for the SPTD and 18Mbps for the baseline. The 5-percentile comparison shows a bigger gap with 17.6Mbps for MPTD, 16.4Mbps for SPTD and 13.8 Mbps for the baseline, leading to a better usage efficiency of the air interface with either SPTD or MPTD.

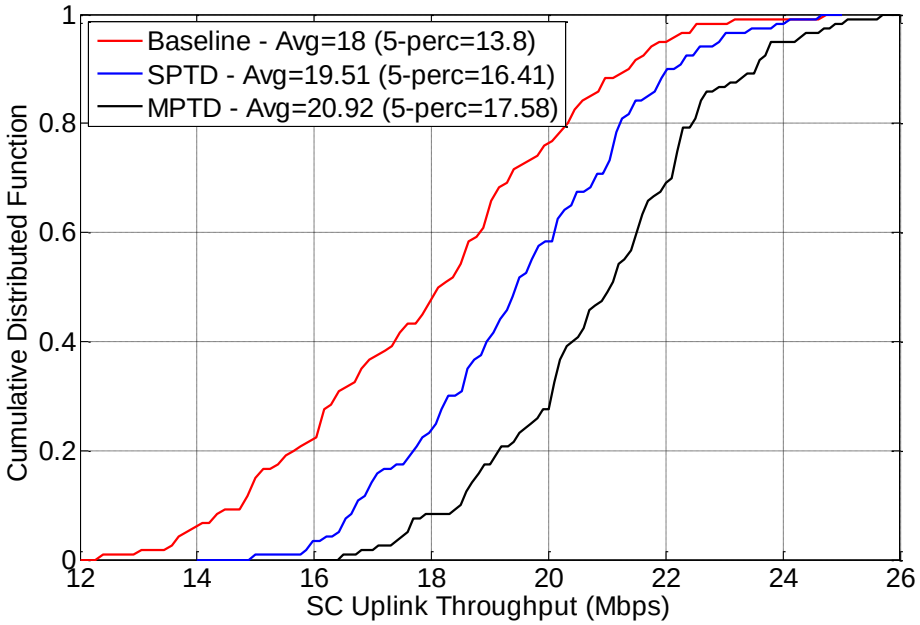


Figure 4-44: Comparison of the small-cell uplink throughput CDF

The previous figure also shows that the centralised RRM algorithm improves the area throughput. On a per user basis, the gain is also in favour of the SPTD/MPTD solution compared to the baseline as depicted by Figure 4-45. For instance 30% gain can be observed for MPTD for the 5-percentile value, which represents the edge user throughput.

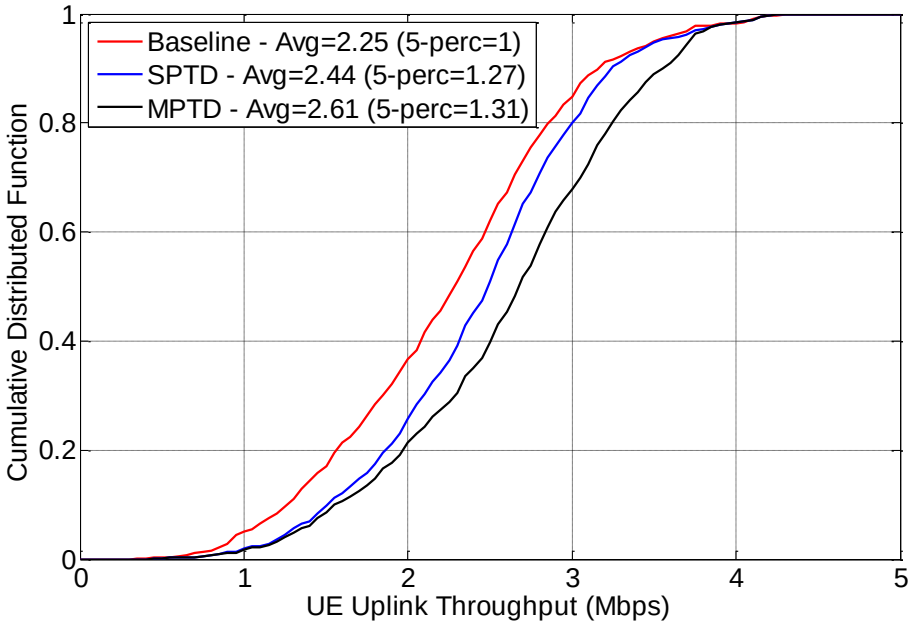


Figure 4-45: Comparison of the user uplink throughput CDF

Since the centralised RRM algorithm only relies on the downlink RSRP measurements and not an exact knowledge of the uplink conditions (channel state information for instance), it is interesting to compare the throughput of the users which would have been paired together in the baseline scenario against the throughput of the SPTD/MPTD users. Figure 4-46 shows such comparison, where the term "MPTD candidate" designates the paired users. On a side note the "paired" users in the baseline scenario are of

course not scheduled on the same resource as it would introduce a bias on the comparison. They are SPTD/MPTD candidates in the sense that the centralised algorithm would have paired them together.

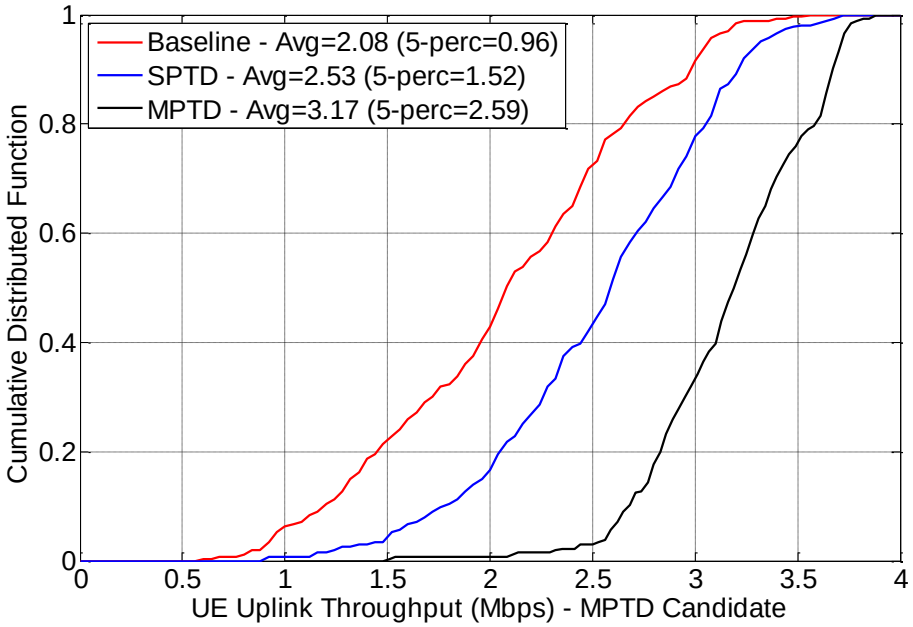


Figure 4-46: Comparison of the "paired" user uplink throughput CDF

We can notice that the MPTD really brings a significant advantage to the "paired" users, with a throughput which is on average 50% greater than the baseline (3.2Mbps versus 2.1Mbps). But more importantly, the 5-percentile shows a bigger gap (2.6Mbps versus 1Mbps) thanks to the centralised processing leading to a "fairer" situation among the paired users.

The results have shown that MPTD or even SPTD without cooperation during the physical processing can lead to a higher area throughput and utilisation efficiency (at least of the air interface). In the next steps, we will investigate the backhaul constraint and how we may circumvent them as well as a more "dynamic" centralised RRM algorithm where more metrics will be used to determine if a user may be a candidate for the pairing (not just the RSRP).

4.8 CT 3.8: Radio Resource Management for In-Network-Processing

4.8.1 Technical description

Scenario

In-Network Processing (INP) is a technique that allows for a distributed Multi-User Detection (MUD) of UE uplink signals over several iSCs and can therefore be regarded as a CoMP technique. Compared to centralized CoMP, where the joint processing takes place in the RANaaS, the received signals are not forwarded directly to the RANaaS, causing large traffic on the J1 link, but the iSCs exchange information among each other over J2 links [14].

System model

This candidate technology is the RRM counterpart of WP2 CT2.1, considering the actual detection algorithms, described in detail in the deliverable D2.2 [43]. In this CT, we will evaluate the algorithms developed within CT2.1 for several resource allocation examples based on standard scheduling approaches.

Approach

The use of MUD allows UEs to transmit data on the same physical resources, i.e., a frequency reuse of 1 among neighboring /overlapping iSCs. The RRM needs to consider this possibility and should schedule UEs on the same PRBs if a) the channel promises a good separability, b) sufficient J2 backhaul capacity is available to facilitate the exchange of information among iSCs and c) the added processing latency can be tolerated.

4.8.2 Implementation of CT in the iJOIN architecture

The INP-aware multi-iSC RRM can either be implemented at one iSC or at the RANaaS, as proposed in the deliverable D3.1 [5]. In the case that the RRM is implemented at the RANaaS, which in general is preferred, the information exchange described by the message sequence chart in Figure 4-47 is observed. For every exchanged parameter, the corresponding identifier (I3.x, O3.x) is given as defined in D3.1 [5] and summarized in Appendix I.

Compared to a conventional scheduler, the INP-aware scheduler collects information from several iSCs as well as backhaul information obtained by the iNC in order to decide which UEs are allocated onto the same physical resources and thus need to be detected jointly among iSCs.

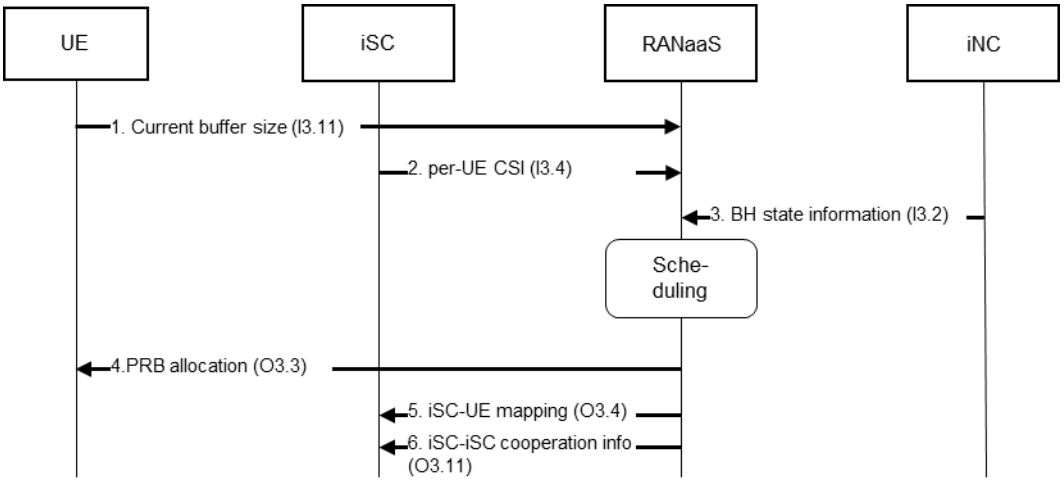


Figure 4-47: Message sequence chart for INP-aware central RRM as proposed by CT3.8

4.8.3 Evaluation of the CT

Performance evaluation is done by means of numerical simulations. In particular, a link level simulator has been developed for the WP2 simulation activities. This link level simulator will be used in combination with exemplary RRM allocations based on the common evaluation scenarios.

Compliance with iJOIN objectives

The scheduling of several UEs to the same physical resources allows for an increase in area throughput. If INP is applied to uplink signals of orthogonally scheduled UEs, an SNR gain can be achieved, allowing for a reduction of UE transmit power, improving a kind of energy efficiency w.r.t. UE energy. Additionally, by distributing the detection over several iSCs in the reception range, taking into account the current backhaul load an improvement of UEff can be achieved.

Description of the baseline used for the evaluation

The baseline is a LTE Release 10 environment with orthogonal PRB allocation, which will always be simulated alongside.

The output of the link level simulations will be BER/FER curves for different MCS and different allocations (orthogonal / non-orthogonal), from which the achievable spectral efficiency and thus, the area throughput can be extrapolated.

Generally, the scenarios addressed are the outdoor square scenario and the indoor scenario. Initially, a simple, academic scenario will be used, but subsequently, more parameters (e.g., channel model) as defined in appendix II.2 will be introduced.

As a toy example, which has also been used in WP2 CT2.1, a simple setup consisting of 4 iSCs will be used. Three different exemplary backhaul topologies are depicted in Figure 4-48, a network with Point-to-Point links, a network with wireless Point-to-Multipoint backhaul and a configuration with Point-to-Point links where an iTN serves as a switching node. In [42] and [53], these topologies have been compared w.r.t. the resulting backhaul load.

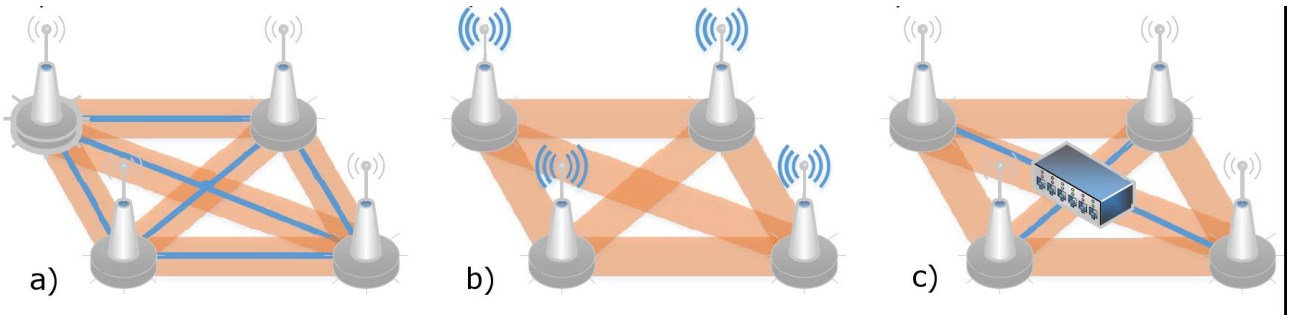


Figure 4-48: Exemplary physical backhaul topologies for 4 iSCs, a) Point-to-Point, b) Point-to-Multi-Point, c) Point-to-Point with central iTN

Discussion of results of the CT

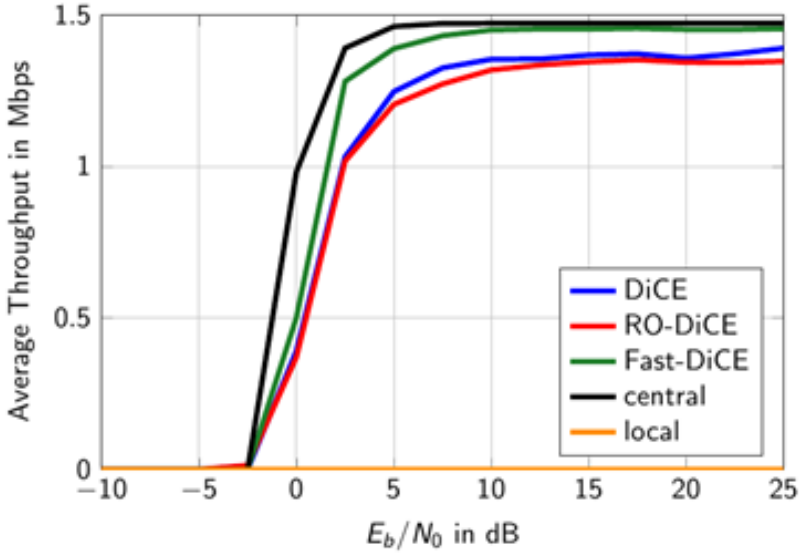


Figure 4-49: Average throughput for different INP variants and central processing

Link level simulations have been performed of a system with 1.4 MHz bandwidth where 4 UEs were allocated the same physical resources, with each UE using LTE MCS 7 and 2 spatial streams. MUD was performed by 6 iSCs in a logical full mesh topology, using 2 receive antennas each. Figure 4-49 shows the average throughput for 10 iterations of 3 different variants of the Distributed Consensus Based Estimation (DiCE) INP algorithm; the original DiCE, the Reduced Overhead DiCE (RO-DiCE) and the Fast-DiCE [43], compared to centralized and local processing. It can be seen that almost the same throughput as for centralized processing can be achieved, at heavily reduced J1 load, however, at the expense of significant J2 load.

Further results for the reference scenarios using different resource allocations (orthogonal, partial overlap, full overlap) will be provided in the final deliverable, where also the trade-off between area throughput and backhaul load will be addressed.

4.9 CT 3.9: Hybrid local-cloud-based user scheduling for interference control

4.9.1 Technical description

Scenario

Joint cooperative scheduling allows significant performance improvements. It comes however with strong requirements in terms of exchange of information. One major challenge is to obtain the performance improvement of joint scheduling with only limited exchange of information between the cooperating transmitters. Hence, the goal of this CT is to develop new cooperative scheduling algorithms which efficiently exploit any backhaul topology available. This question is especially challenging because it is not

clear upfront what kind of information should be exchanged between the iSCs and the RANaaS and how to split the scheduling between the RANaaS and the iSCs.

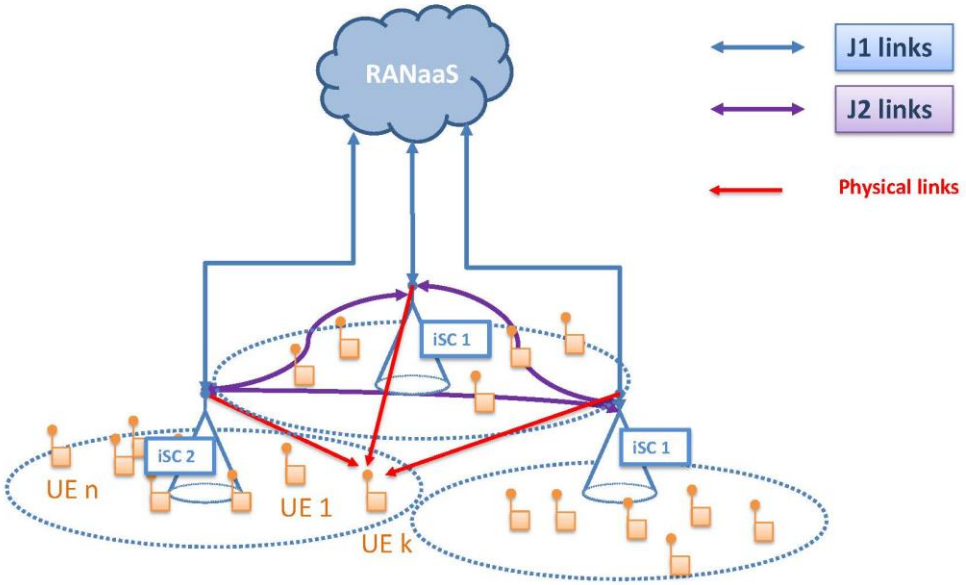


Figure 4-50 Schematic presentation of the architecture as studied in CT3.9

In particular, we propose in this deliverable a distributed scheduling scheme relying on a Bayesian analysis performed at each iSC to maximize the expected sum rate.

In the scenario considered, the scheduling is done at each iSC on the basis of only the instantaneous knowledge of the direct channel to its own user and the long terms information of the multi-user channel. This scenario models the fact that sufficient feedback resources are available to obtain the knowledge of the direct channel, but the backhaul links are weak or suffer from an important delay such that it is only possible for the iSCs to exchange in the backhaul the long term information.

System model

We focus then in this CT on a setting with K iSCs serving K single-antenna UEs. We denote the channel gain from the ith iSC to the kth UE by $G_{k,i}$. The transmission scenario is illustrated in the case of $K = 2$ in Figure 4-51.

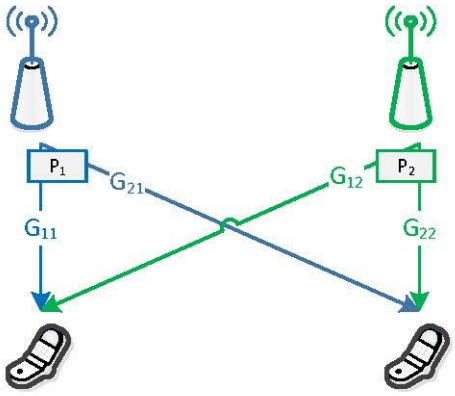


Figure 4-51: Cooperative power control with distributed CSI at the iSCs.

In that setting, we aim then at finding the scheduling, or equivalently the binary power control, which maximizes the ergodic sum rate defined as:

$$\mathrm{E}[\mathrm{R}(P_1,\ldots,P_K)] = \mathrm{E}[\sum_{k=1}^{K} \mathrm{R}_k(P_1,\ldots,P_K)] \tag{4.56}$$

where

$$\mathrm{R}_k(P_1,\ldots,P_K)=\log_2\left(1+\frac{P_k G_{k,k}}{1+\sum_{j=1,\,j\neq k}^{K} G_{k,j}P_j}\right). \tag{4.57}$$

We consider distributed scheduling such that each iSC has to decide for the given realization of the received channel gain $G_{j,j}$ whether to transmit or not.

With full centralized CSI, the optimal scheduling solution is easily obtained by simply computing for each case the sum rate, and choosing the one which leads to the largest sum rate. Note however that if this solution can easily be applied only when the number of iSCs is small as the number of possibilities to test increases exponentially with the number of iSCs. With only local CSI knowledge, finding the optimal scheduling decision is a difficult problem as it falls in the category of Team Decision problems [39]: The iSCs aim at jointly maximizing a common objective on the basis of individual information.

In fact, the distributed Team Decision problem can be reformulated as a conventional optimization problem as follows:

$$(p_1^*,\ldots,p_K^*)=\underset{(p_1\ldots,p_K)}{\arg\max}\,E[R(p_1(G_{1,1}),\ldots,p_K(G_{K,K}))] \tag{4-58}$$

Where

$$p_j : G_{j,j} \mapsto p_j(G_{j,j}). \tag{4-59}$$

It is important to note that with this reformulation, the optimization variables are no longer scheduling decision but *scheduling functions* with the appropriate dependencies.

This optimization being in general very difficult to solve, the usual approach consists in studying instead the best-responses functions [40]. This comes down to finding the scheduling functions verifying

$$p_j^{BR}=\underset{p_j}{\arg\max}\,E[R(p_1^{BR}(G_{1,1}),\ldots,p_{j-1}^{BR}(G_{j-1,j-1}),p_j(G_{j,j}),p_{j+1}^{BR}(G_{j+1,j+1}),\ldots,p_K^{BR}(G_{K,K}))]\ . \tag{4-60}$$

In words, this means finding the scheduling functions which are optimal for each iSC, given the scheduling functions of the other iSCs.

Approach

One of the main difficulties in solving optimization problem (4-60) comes from the fact that the optimization is done over a functional space of infinite dimension. However, it can be easily shown from the monotonic behaviour of the sum rate with respect to the gain of the direct channels, that the optimal scheduling functions are *thresholds functions*, i.e. can be written as [38].

$$p_j^{\lambda_j}(G_{j,j})=\begin{cases}0 & \text{if } G_{j,j}\le\lambda_j\\ P_j^{\max} & \text{if } G_{j,j}>\lambda_j\end{cases}. \tag{4-61}$$

Hence, the functional optimization (4-60) problem can be reformulated as the finite dimensional optimization problem

$$\lambda_j^{BR}=\underset{\lambda_j}{\arg\max}\,E[R(p_1^{\lambda_1^{BR}}(G_{1,1}),\ldots,p_{j-1}^{\lambda_{j-1}^{BR}}(G_{j-1,j-1}),p_j^{\lambda_j}(G_{j,j}),p_{j+1}^{\lambda_{j+1}^{BR}}(G_{j+1,j+1}),\ldots,p_K^{\lambda_K^{BR}}(G_{K,K}))]\ . \tag{4-62}$$

This optimization still requires evaluating the expectation over the channel coefficients. In general, this expectation is difficult to tackle in closed form and we will use Monte-Carlo simulations to approximate its value.

The best-response update can be done without a large complexity as it can be easily obtained that the best-response value can be obtained using the well-known and very efficient bisection algorithm. More details on the best-response algorithm to iteratively update the thresholds can be found in [38].

4.9.2 Implementation in the iJOIN architecture

The goal of this CT is to exploit in an optimal (Bayesian) manner the information which is locally available so as to perform the most efficient distributed scheduling decision possible. Hence, it requires a functional split where the scheduling decision is done at least partially at the iSCs.

This CT requires the downlink CSI relative to the direct channel gain to be available at the iSC. Furthermore, the optimization to obtain the scheduling functions is done on the basis of the statistics of the multi-user channel. Hence, it requires the knowledge of these statistics either at every iSC to do the computing in a distributed manner at the iSCs or at the RANaaS where it will be done centrally.

Therefore, either J1 links between every iSC and the RANaaS have to be available or J2 links between iSC. In both cases, the information exchange need to occur only over a long term with respect to the channel statistics, hence in the order of several seconds.

This exchange of messages in the iJOIN architecture is shown in Figure 4-52.

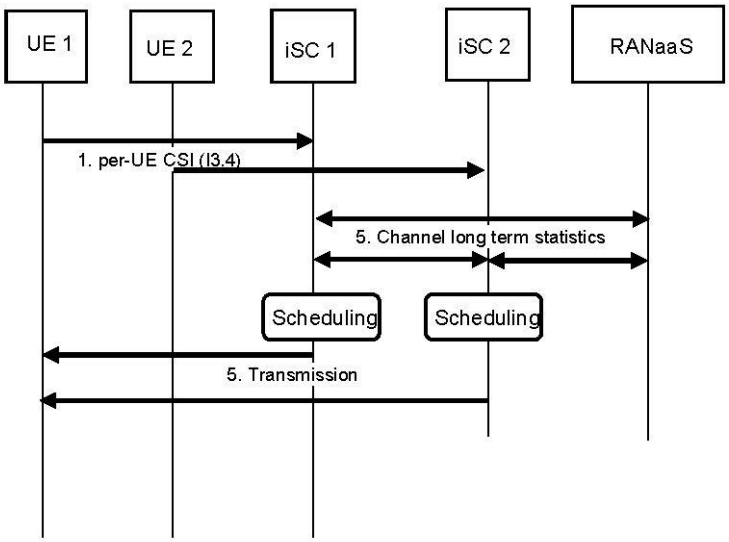


Figure 4-52: Message sequence chart for the hybrid scheduling algorithm as proposed by CT3.9

4.9.3 Evaluation of the CT

Compliance with iJOIN objective

This CT aims at adapting to backhaul links of varying qualities and hence achieving a flexible interference management. It hence deals with one of the core issues of the iJOIN approach which is flexible cooperative scheduling. The main focus is the analysis of the average performance such that this CT focuses on the area throughput.

Description of the baseline used for the evaluation

We compare our approach to three schemes. The first one corresponds to the case of infinite backhaul where it is possible to do a perfect joint scheduling and provides hence an upper bound to the performance. The second scheme is the non-cooperative one where each iSC emits with full power. Finally, the last scheme of comparison is the Round-Robin where perfect coordination is achieved but there is not opportunistic gain. Both alternatives correspond to what would be done in an LTE network with only knowledge of the downlink CSI.

Discussion of results of the CT

We use Monte-Carlo simulations with 10000 realizations. We further consider in the simulations a Rayleigh fading environment with channel gains being i.i.d. with the variance profile $\sigma_{11}^2=1,\sigma_{12}^2=1,\sigma_{21}^2=1,\sigma_{22}^2=1$. The simulation results are shown in Figure 4-53.

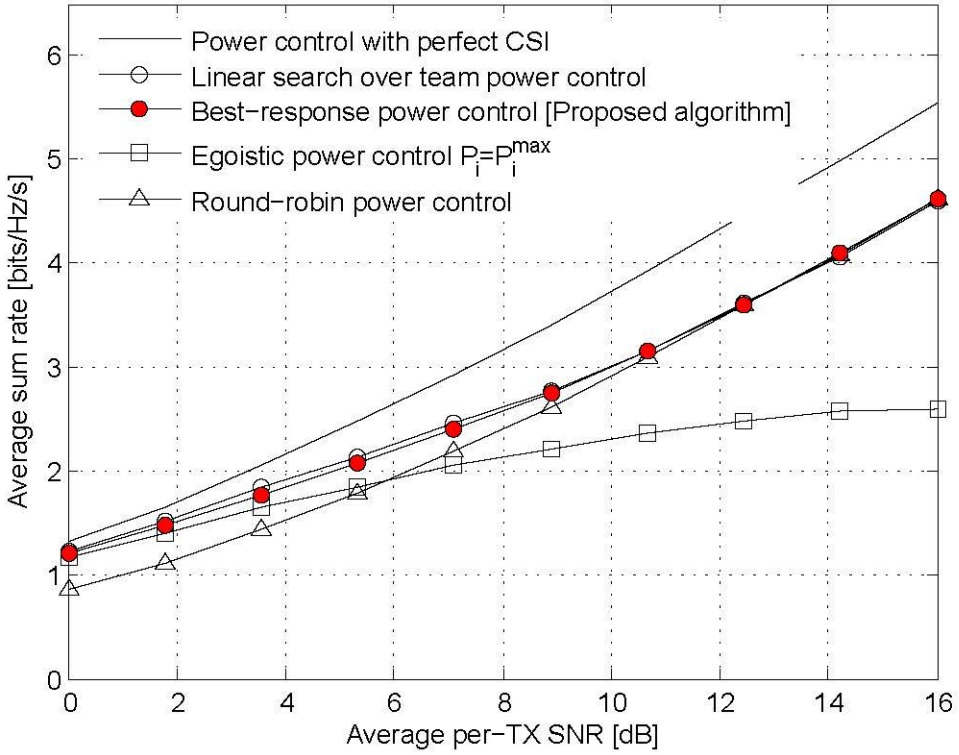


Figure 4-53: Ergodic rate achieved with the different scheduling strategies for the pathloss parameters $\sigma_{11}^2=1,\sigma_{12}^2=1,\sigma_{21}^2=1,\sigma_{22}^2=1$.

It can be seen that the proposed Bayesian scheduling approach outperforms both conventional scheduling methods and goes smoothly from one to another. This behaviour is intuitively meaningful as it can be easily shown that the egoistic scheduling is optimal at low SNR while round-robin becomes optimal at large SNR. We also compare the proposed best-response approach to a line search of the threshold being optimal for the initial Team Decision optimization problem described in equation (4-62).

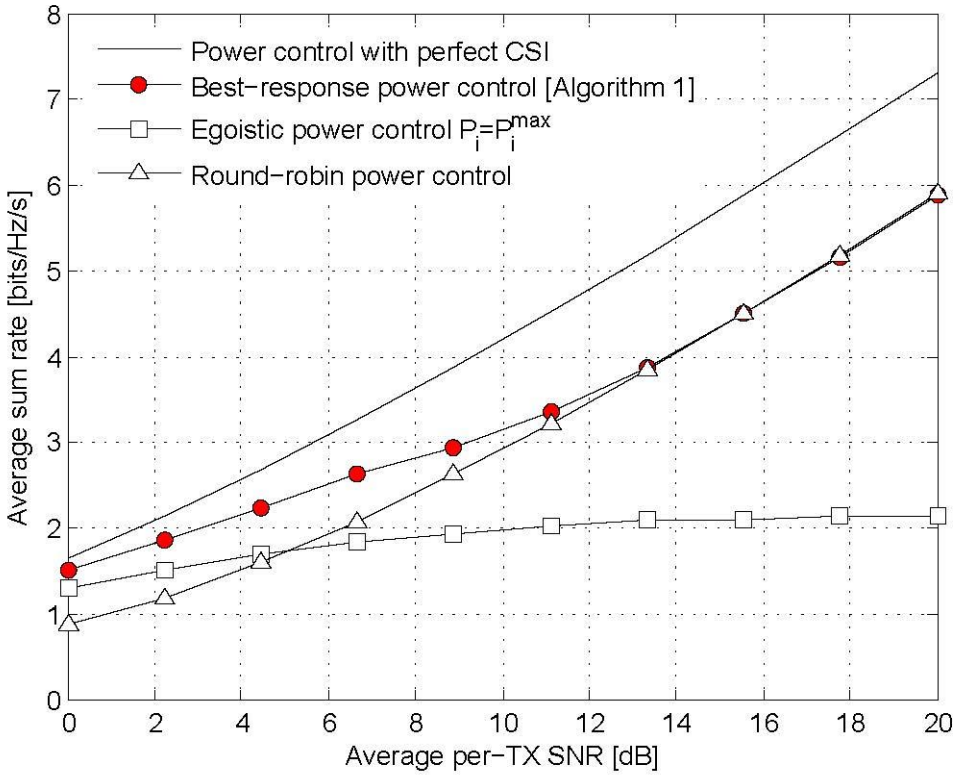


Figure 4-54: Ergodic rate achieved with the different scheduling strategies for uniform pathloss with $K = 5$.

Finally we show in Figure 4-54 the sum rate achieved also with uniform pathloss but in the case of $K = 5$ in order to have a first evaluation of the scaling behaviour in term of the number of iSCs. In fact, it has be shown in [41] in the case of i.i.d. coefficients that the proposed distributed scheduling approach achieves the same scaling law in the number of iSCs as the optimal centralized scheduling on the basis of perfect CSI. Hence, the proposed distributed scheduling approach appears as an interesting low complexity method to achieve most of the gains due to coordinated scheduling at a low cost in terms of backhaul resources and computation resources.

This asymptotically optimal behaviour can be observed in the figure where the convergence of the performance obtained with the proposed algorithm to the performance obtained using Round-Robin is observed at higher SNR than in the two-user case.

5 Overall Evaluation

In this Section, we aim to obtain a preliminary comparison of the overall WP3 CTs as well as to define possible configurations of compatible CTs, which can be integrated in order to achieve the global iJOIN targets.

First, we summarize the qualitative impact of the CTs towards the iJOIN objectives: Area Throughput (AT), Energy Efficiency (EEff), Cost Efficiency (CEff), and Utilization Efficiency (UEff). In Table 5-1 the "++" symbol indicates that a given CT mainly affects positively a specific objective, "+" accounts for beneficial side effect and the "0" represents a negligible impact. As expected most of the WP3 CTs target to improve the AT; however, a high number of CTs also ameliorate the system UE by enhancing the effectiveness in the resource usage.

Note that some CTs can be based on a specific fundamental trade-off; therefore, it is important to set up the CTs such that it does not negatively affect the iJOIN objectives. For instance, improving AT by using a joint transmission scheduler, where multiple iSCs jointly send a given message to a single UEff, may require an increase of the overall energy consumption. In this case, the number of cooperating iSCs has to be limited such that the overall EEff (the ratio between the energy consumption and the provided) is not reduced.

Table 5-1: Qualitative impact of the WP3 CTs with respect to the global iJOIN objectives.

	AT	**EEff**	**CEff**	**UEff**
CT 3.1	++	0	0	+
CT 3.2	++	0	0	+
CT 3.3	0	++	+	0
CT 3.4	++	0	+	0
CT 3.5	++	0	0	+
CT 3.7	++	0	0	+
CT 3.8	++	0	0	+
CT 3.9	++	0	0	+

In Table 5-2 we present the relationships between CTs in terms of compatibility. Compatible CTs can be successfully integrated to jointly enhance the performance of a reference system. This integration may require some kind of coordination between CTs that operate on the same resources but on a different time scale, e.g. inter-cell interference coordination and scheduling algorithm. On the contrary, CTs that conflict with each other cannot be used jointly since they simultaneously operate e.g. on the same resource or functionality.

WP3 CTs are characterized by a limited number of possible conflicts:

- CT3.4 "Computational Complexity and Semi-Deterministic Scheduling", CT3.5 "Computation complexity and semi-deterministic scheduling", and CT3.9 "Hybrid local-cloud-based user scheduling for interference control" cannot be implemented jointly since they all focus on downlink radio resource management.
- CT3.7 "Radio resource management for scalable multi-point turbo detection/In-network Processing" may not be compatible with CT3.8 "Radio Resource Management for In-Network-Processing", which is a concurrent uplink RRM method.

Table 5-2: Compatibility of the WP3 CTs.

	CT 3.2	CT 3.3	CT 3.4	CT 3.5	CT 3.7	CT 3.8	CT 3.9
CT 3.1	Y	Y	Y	Y	Y	Y	Y
CT 3.2		Y	Y	Y	Y	Y	Y
CT 3.3			Y	Y	Y	Y	Y
CT 3.4				N	Y	Y	N
CT 3.5					Y	Y	N
CT 3.7						N	Y
CT 3.8							Y

Definitions:

"**Y**" – interoperable: CTs operate on different resources or in different operational domains. Algorithmic coordination and/or information exchange with iJOIN network entities (e.g. iNC or iveC) may be necessary. Different domains/resources include:

- Backhaul:
  - Channel resources (e.g. wireless, wired, …)
  - Link/Routing
- RAN:
  - RF transmission (transmit power)
  - Downlink radio resources
  - Uplink radio resources
  - Cell association

"**N**" – not interoperable: CTs operate on same resources and/or are based on different assumptions

This work will be continued in deliverable D3.3, which focuses on the final definition and evaluation of MAC and RRM approaches for RANaaS and a joint backhaul/access design. There, the evaluation of the CTs under the pre-defined iJOIN common scenarios will be further discussed. Moreover, a comparative study will capture the aggregated gains in the aforementioned objectives, where sets of compatible CTs are combined under the same framework. In this direction, we provide one table per CT in Appendix IV, where we show the interactions between different CTs. These tables give a more detailed description of the CTs' compatibility and are used as inputs in order to capture the CT interactions and prepare Table 5-2. The effort to show the compatibility between different CTs will be further discussed and finalized in D3.3.

6 Summary and Conclusion

In this report, the concept of the veNB was readdressed and its actual implications on the proposed CTs were investigated. In particular, the key concept of a functional split between local and centralized processing was looked at from the virtual eNB perspective. In this context, the key layer 2 functions were highlighted and the functional split was further de-composed to capture the effect of centralization for different functions. Different functional split options have been evaluated according to their benefits and drawbacks in terms of potential centralization gains, implementation and architectural impact as well as backhaul requirements. Furthermore, key decision factors were defined to help us identify which should be the best functional split configuration to be selected in a specific network deployment. Moreover, the veNB implementation aspects were further described, incorporating the per-CT virtualization and some functional constraints imposed by the RANaaS platform.

Furthermore, we have refined the set of candidate technologies introduced in deliverable D3.1 [5] and outlined their applicability to the iJOIN architecture defined in deliverable D5.1 [15]. By means of initial evaluation, it was depicted how the iJOIN key objectives are addressed by the proposed approaches, and how practical constraints and requirements affect the different approaches.

In order allow the comparison of results from the different CTs using different approaches, we provided a CT compatibility study which shows which CTs are operationally compatible in the iJOIN system model. In this direction, a set of common simulation parameters for the evaluation of the CTs has been defined. 2 models, one outdoor and one indoor, are defined and can be mapped to the common scenarios defined in D5.1 [15]. Subsequently, the categorization of backhaul technologies was also provided, showing some key measures (latency, throughput) that can be used as guidelines for the CT evaluation and the functional split selection.

Acknowledgements and Disclaimer

This work was partially funded by the European Commission within the 7th Framework Program in the context of the ICT project iJOIN (Grant Agreement No. 317941).

Appendix I Input and Output Parameters

Based on the functional architecture defined in deliverable D5.1 [15], the input and output interfaces for WP2 CTs were defined in deliverable D3.1 [5]. Table 6-1 and Table 6-2 describe the input and output information required by each CT. They list the related CTs, requested input information or provided output information, the sink or source of information in terms of CT and logical network entity, and lists the parameterization of the interface. Furthermore, Table 6-3 describes each acronym and we indicate whether the related parameter is already defined in 3GPP LTE or has been introduced by iJOIN.

Table 6-1: Required Input of WP3 CTs

IP	CT	Requested Input	Source of Information		
			CT or system function	Logical network entity	Parameters
I3.1	3.1, 3.5 3.8	BH Routing Table /Info	C.T. 4.4	iNC	<src_address> <gateway> (nexthop) <dst_address>
I3.2	3.1, 3.2 3.3 3.8 3.9 3.7 3.5 3.4	BH state information (iSC-iSC, iSC-iTN) • SINR • Max Capacity • Remaining Capacity	Measurements iTN	iTN	<BH_SINR>, <BH_MAX_CAP>, <BH_RES_CAP>
			Measurements report iNC	iNC	<BH_ID>, <BH_SINR>, <BH_MAX_CAP>, <BH_RES_CAP>
I3.3	3.1, 3.2, 3.3, 3.4, 3.5, 3.7	QoS parameters per bearer (e.g. max. bit rate (MBR), guaranteed bit rate (GBR), packet delay budget (PDB))	RRC	veNB	<MBR>, <GBR>, <PDB>
I3.4	3.2, 3.3, 3.4, 3.5, 3.7, 3.8, 3.9	Channel state information per UE (DL/UL) (CQI,RSRP,..)	RRC (measurements)	veNB	<WCQI>, <SCQI>, <PMI>, <PTI>, <RI>, <RSRP>, <RSRQ>
I3.7	3.3, 3.4, 3.5, 3.7	Cell ID (to which cell the UE is currently connected, or which is the current location of the UE (depending on the RRC/ECM state))	In RRC_CONNECTED state: RRC (ECGI) In RRC_IDLE state: MME (last known ECGI)	veNB/MME	<ECGI>
I3.10	3.2	Scheduling policy (i.e., MCI, EDF,PF)	Scheduling	NMS	<SCHED_POL>
I3.11	3.3, 3.5, 3.9	Current buffer size • DL • UL	MAC (BSR)	veNB	<UL_BSR>, <DL_BSR>
I3.12	3.4, 3.7	RNTI	RRC	veNB	<RNTI>
I3.13	3.4	iSC → RANaaS (J1): Quantized CSI	RRC	iSC	<BH_ID>, <BH_QUANT_CSI>
I3.14	3.4 3.5	iSC ←→ iSC (J2): Interference coordination information	CT 3.4	iSC	<RNTP>
I3.16	3.5, 3.7	Latency backhaul (iSC->iSC; RaaS-iSC)	Measurements iTN	iNC	<BH_LAT>
			Measurements report iNC	iTN	<BH_ID>, <BH_LAT>

IP	CT	Requested Input	Source of Information		
			CT or system function	Logical network entity	Parameters
I3.18	3.7	UE mobility state information	RRC	veNB	<UE_MSE>
I3.21	3.7	UE capability (category …)		veNB/MME	<UE_CAP>
I3.22	3.2	Cell neighbouring list	RRC or OAM	veNB or SON	Array of <ECGI>

Table 6-2: Required Output of WP3 CTs

OP	CT	Provided Output	Sink of Information		
			CT or system function	Logical network entity	Parameter
O3.1	3.1	BH Link Selection	Scheduling	iTN	<BH_ID>
O3.2	3.1, 3.3, 3.9	RRM information (allocation of resources per BH link)	Scheduling	iTN	<BW>, <BH_FREQ>
O3.3	3.2, 3.3, 3.4, 3.5, 3.7, 3.8, 3.9	Resource allocation per UE (RBs, …)	Resource Mapper (e.g. in CT 2.1)	iSC	<SFN>, <DCI>
O3.4	3.2 3.8	iSC-UE mapping	CT2.3.2.1/RRC	veNB	<UE_ID>,array of <ECGI>
O3.5	3.3, 3.4, 3.5, 3.9	MCS (access)	T2.3 Scheduling	iSC	<MCS>
O3.6	3.4	RANaaS → iSC (J1): Long term/coarse grained resource schedule	Scheduling	iSC	<LT_SCHED>
O3.7	3.5	Cell DTX pattern (for both U/C planes)	Scheduling	veNB	<DTX_PATTERN_ID>
O3.8	3.7	Local or Centralised computing	CT2.2	iSC	<SPLIT_CONF>
O3.9	3.7	Per iSC: • pair of (iSC-UE) involved in the turbo processing	CT2.2	veNB	Array of pair (<iSC_ID>, <UE_ID>)
O3.11	3.8	List of cooperating iSCs per iSC with max. bandwidth	CT2.1	iSC	Array of <iSC_ID>-<BH_MAX_CAP > tuples per iSC

Table 6-3: List of Abbreviations

Abbrev	Full Name (including explanation if necessary)	LTE or CT specific
Identifier		
BH_ID	Backhaul link identifier (for logical network)	iJOIN
BH_SINR	Backhaul link SINR	iJOIN
BH_MAX_CAP	Maximum backhaul link capacity (kbps)	iJOIN
BH_RES_CAP	Residual backhaul link capacity (averaged, kbps)	iJOIN
ECGI	cell ID	LTE
UE_ID	Unique identifier of an UE	iJOIN
MBR	Maximum Bit Rate (bps)	LTE
GBR	Guaranteed Bit Rate (bps)	LTE
PDB	Packet Delay Budget (ms)	LTE
WCQI	Wideband CQI	LTE
SCQI	Subband CQI	LTE
PMI	Index of precoding matrix (TS 36.213)	LTE
PTI	Index of precoding type (TS 36.213)	LTE
RI	Rank indication (TS 36.213)	LTE
SCHED_POL	Scheduling policy in iSC (e.g. RR, PropFair)	iJOIN
UL_BSR	Uplink buffer state report	LTE
DL_BSR	Downlink buffer state report	iJOIN
RNTI	Radio Network Temporary Identifier	LTE
BH_QUANT_CSI	Quantized backhaul CSI	iJOIN
RNTP	Relative Narrowband Transmit Power (TS 36.413)	LTE
BH_LAT	Backhaul link latency (ms)	iJOIN
UE_MSE	UE mobility state (depending on velocity) (high, medium, low)	LTE
UE_CAP	UE capability	LTE
SFN	System Frame Number	LTE
DCI	Downlink Control Information	LTE
UCI	Uplink Control Information	LTE
MCS	Modulation and coding scheme for access	iJOIN
DTX_PATTERN_ID	ID of cell DTX pattern	iJOIN
LT_SCHED	Long term/coarse grain resource schedule (FFS)	iJOIN
SPLIT_CONF	Indicates functional split configuration to CT	iJOIN

Appendix II Evaluation methodology

This section captures the agreements of the iJOIN WP3 partners on common reference system parameters. These results from intense coordination work are an indispensable prerequisite for a solid and comparable quantification of the solutions proposed in iJOIN WP3. The parameters and settings described here are mainly based on the 3GPP LTE system, which is used in iJOIN as reference point.

The reference scenarios specify deployment and operation assumptions for macro sites and small cells deployed in both outdoor and indoor use cases.

II.1 Radio Access Network Modelling

II.1.1 Outdoor small cell deployment

In this model, outdoor small cells are mainly deployed to create local hotspots in the macro cell area.

In this case, both the macro cell and the small cell may operate on a common band, i.e., co-channel deployment at 2 GHz, or a dedicated band at higher frequency (3.5 GHz) can be used by the small cell to exploit the larger available bandwidth and avoid cross-tier interference (see Figure 6-1).

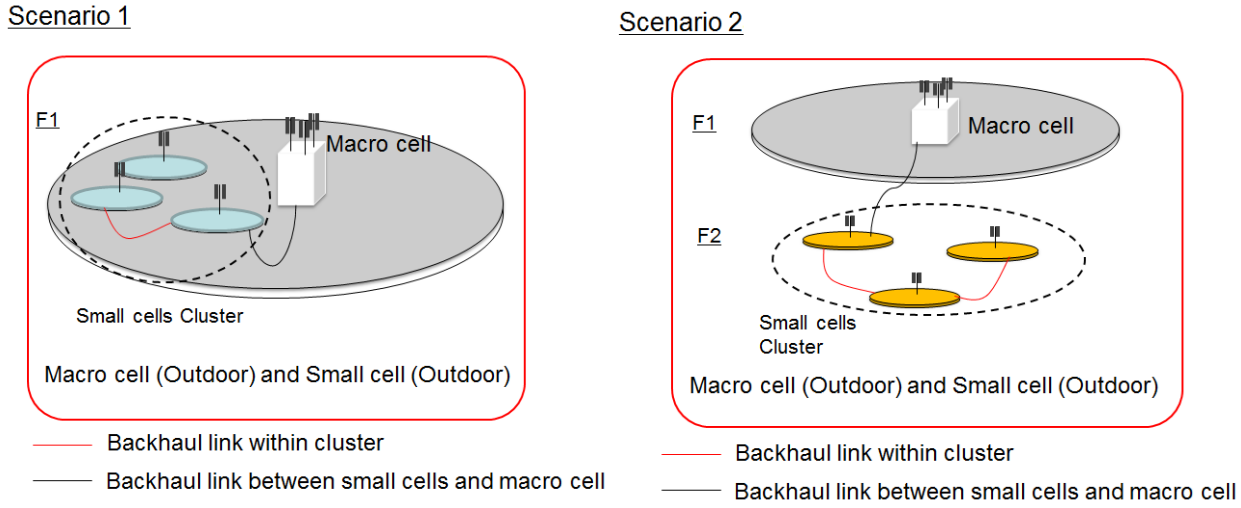


Figure 6-1: Deployment scenarios of outdoor small cells with macro coverage [21].

However, following the 3GPP [3] scenarios, we include also cases where the macro cell coverage is disregarded and only small cells are considered (see Figure 6-2).

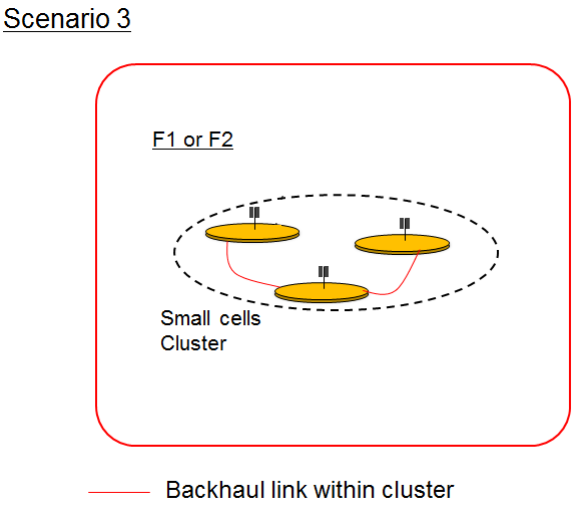
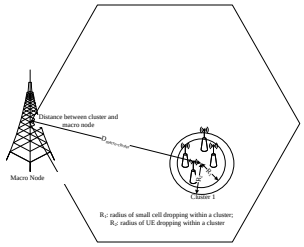


Figure 6-2: Deployment scenarios of outdoor smalls cell without macro coverage.

In both co-channel and dedicated deployments, the density of small cells and end-users can be varied to represent different use case scenarios, for example cluster of small cells can be modelled as well as sparse

hot spots deployment. However, only outdoor UEs are investigated here. Both ideal backhaul and non-ideal backhaul are considered between neighbouring iSCs and between the iSCs and the nearby eNB.

Table 6-4: 3GPP outdoor deployment assumptions [21].

	Macro cell parameter	Small cell parameter
Layout	Hexagonal grid, 3 sectors per site, 19 Macro sites and 7 Macro sites can be used.	 Clusters uniformly random within macro geographical area; small cells uniformly random dropping within cluster area
System bandwidth per carrier	10MHz	10MHz
Carrier frequency	2.0GHz	2.0GHz / 3.5 GHz
Carrier number	1	1
Total BS TX power (Ptotal per carrier)	46dBm	30 dBm, Optional: 24dBm, 37dBm
Distance-dependent path loss	ITU UMa (3GPP TR36.814)	ITU Umi (3GPP TR36.814)
Penetration Loss for indoor UEs	For indoor UEs: 20dB+0.5d_{in} (d_{in} : independent uniform random value between [0, min(25,d)] for each link)	For indoor UEs: 20dB+0.5d_{in} (d_{in} : independent uniform random value between [0, min(25,UE-to-eNB distance)] for each link)
Shadowing	ITU UMa (3GPP TR36.819)	ITU UMi (3GPP TR36.814)
Antenna Height:	25m	10m
UE antenna Height	1.5m	
Antenna gain + connector loss	17 dBi	5 dBi
Antenna gain of UE	0 dBi	
Fast fading channel between eNB and UE	ITU UMa (3GPP TR36.814)	ITU Umi (3GPP TR36.814)
Antenna configuration	2Tx2Rx in DL, Cross-polarized	
Number of clusters/buildings per macro cell geographical area	1, 2, optional of 4	
Number of small cells per cluster	4, 10	
Number of small cells per Macro cell	[4,10]*Number of clusters per macro cell geographical area	
Number of UEs	60 UEs per macro cell geographical area are recommended when FTP model 3 is used	
UE dropping	Baseline: 2/3 UEs randomly and uniformly dropped within the clusters, 1/3 UEs randomly and uniformly dropped throughout the macro geographical area. 20% UEs are outdoor and 80% UEs are indoor.	
Radius for small cell dropping in a cluster	50m	
Radius for UE dropping in a cluster	70m	
Minimum distance (2D	Small cell-small cell: 20m	

distance)	Small cell-UE: 5m	
	Macro –small cell cluster center: 105m	
	Macro – UE : 35m	
	cluster center-cluster center: 2*Radius for small cell dropping in a cluster	
Traffic model	Baseline: FTP Model 1 as in TR 36.814	
UE speed	3km/h	
Cell selection criteria	Baseline: RSRP for intra-frequency and RSRQ for inter-frequency, with cell common bias if CRE is applied.	

Other relevant parameters are presented in Table 6-4: this table resumes the different assumptions proposed in 3GPP TR 36.872 [21] for the evaluation of mechanisms devoted to outdoor small cells. Further details are also presented in 3GPP TR 36.814 [23].

Table 6-5: Mapping of the iJOIN WP3 assumptions to the 3GPP outdoor model.

Parameter	CEA	NEC	UoB	TUD	UNIS	IMC	IMDEA
Layout	☑	☑	Sparse	☑	☑	Sparse	☑
System bandwidth per carrier	☑	☑	☑	☑	☑	☑	☑
Carrier frequency	☑ 3.5 GHz	☑	☑	☑	☑ 3.5 GHz	☑	☑
Carrier number	1 carrier	1 carrier	1 carrier	1 carrier	1 carrier	1 carrier	1 carrier
Total BS TX power	☑	☑	n/a (UL)	☑	☑	☑	☑
Distance-dependent path loss	☑	☑	☑	☑	☑	☑	☑
Number of small cells per cluster	Varying	Varying	Varying	Varying	Varying	Varying	Varying
Number of UEs	Varying	Varying	Varying	Varying	Varying	Varying	Varying
UE dropping	Outdoor only	Outdoor only	Outdoor only	Outdoor Only	Outdoor only	Outdoor only	Outdoor only
Radius for small cell dropping in a cluster	☑	☑	☑	☑	☑	tbd	☑
Radius for UE dropping in a cluster	☑	☑	☑	☑	☑	tbd	☑
Minimum distance	☑	☑	☑	☑	☑	☑	☑
Traffic model	Full Buffer	☑	Full Buffer	tbd	n/a	Full Buffer	Full buffer
Cell selection criteria	RSRP CRE not applied	RSRP	tbd	tbd	tbd	☒	RSRP
Backhaul model	40/80/120 Mbit/s	ideal backhaul	Varying	Limited capacity	High capacity (60GHz)	Very limited	Varying
Other Simulation variables	-	-	-	Backhaul Delay;	QoS requirements	Power constraint	-
Target metric	Throughput; Energy Efficiency	Utilization efficiency	Through-put	Through-put	Throughput , Delay vs. offered load	Through-put	Through-put

Table 6-5 shows the set of relevant common parameters selected by WP3 partners with respect to the outdoor 3GPP simulation assumptions. Key simulation variables are selected to model different network scenarios. Moreover, the number of users and small cells as well as the backhaul capacity will be varied in numerical simulation to represent distinct load situations. Finally, in line with the global iJOIN goals, throughput, energy efficiency and UEff are the main selected evaluation metrics.

II.1.2 Indoor small cell deployment

In this model, indoor small cells are deployed to provide high data rate services to indoor users, which typically experience poor coverage and limited data rate when being served by the macro eNB, due to propagation/penetration losses. iSCs operate on a dedicated carrier and the macro cell presence is not modelled; thus, cross-tier interference is not considered (see Figure 6-3). The density of small cells can be varied to represent different use case scenarios; however, this scenario mainly fits with the iJOIN CS4 (Indoor (Airport / Shopping Mall) [15].

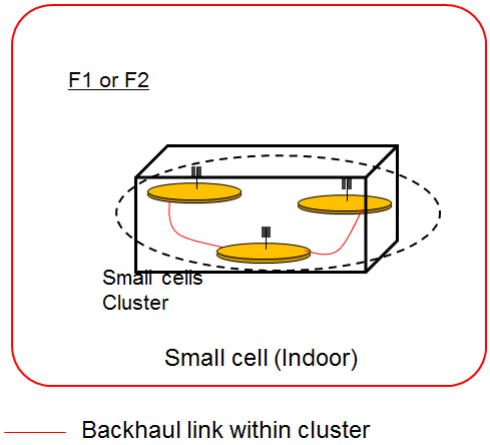


Figure 6-3: Deployment scenarios of indoor small cell without macro coverage [21].

In this scenario, the ITU indoor Hotspot (see Table 6-6) is used in the simulations [21].

Table 6-7 shows the set of relevant common parameters selected by WP3 partners with respect to the indoor ITU/3GPP simulation assumptions. Main simulation variables are selected to model different network scenarios; the number of users and small cells as well as the backhaul characteristics will be varied in numerical simulation to represent distinct load situations. Finally, in line with the global iJOIN goals, throughput is the selected as main evaluation metric.

Table 6-6: ITU indoor deployment assumptions [21].

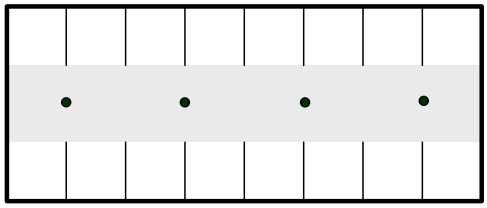
	ITU indoor Hotspot parameters
Layout	 TR36.814; 2/4 small cells per floor, 1/2 floors
System bandwidth per carrier	10MHz
Carrier frequency	3.5GHz
Carrier number	1 or 2
Total BS TX power (Ptotal per carrier)	24dBm
Distance-dependent path loss	ITU InH [referring to Table B.1.2.1-1 in TR36.814]
Penetration	0dB within the same floor;18.3dB between different floors
Shadowing	ITU InH [referring to Table A.2.1.1.5-1 in TR36.814]
Antenna Height:	6m
UE antenna Height	1.5m
Antenna gain + connector loss	5dBi
Antenna gain of UE	0 dBi
Fast fading channel between eNB and UE	ITU InH
Antenna configuration	2Tx2Rx in DL, Cross-polarized
Number of clusters/buildings per macro cell geographical area	N/A
Number of small cells per cluster	N/A
Number of small cells per Macro cell	N/A
Number of UEs	5/10 UEs per small cell
UE dropping	Randomly and uniformly distributed over area per floor
Radius for small cell dropping in a cluster	N/A
Radius for UE dropping in a cluster	N/A
Minimum distance (2D distance)	Small cell-UE: 3m
Traffic model	Baseline: FTP Model 1 as in TR 36.814
UE speed	3km/h
Cell selection criteria	Baseline: RSRP for intra-frequency and RSRQ for inter-frequency, with cell common bias if CRE is applied.

Table 6-7: Mapping of the iJOIN WP3 assumptions to the ITU indoor model.

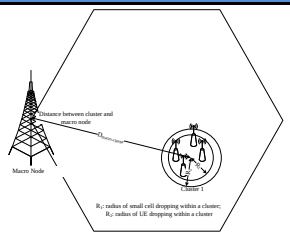
Parameters	SCBB	UoB	UNIS
Layout	☑	☑	☑
System bandwidth per carrier	☑	☑	☑
Carrier frequency	2.6GHz or 3.5GHz	☑	☑
Carrier number	1 carrier	1 carrier	1 carrier
Total BS TX power	☑	n/a (UL)	☑
Distance-dependent path loss	☑	☑	☑
Shadowing	☑	-	☑
Fast fading	☑	-	☑
Antenna configuration	1x2 (UL, ULA)	1x2 (UL)	tbd
Number of small cells per cluster	4	2/4	9-12
Number of UEs	60 (default)	3-5	4-6
UE dropping	Random	Random	Random
Traffic model	Full buffer (other TBD)	Full buffer	n/a
UE speed	☑	0 km/h	n/a
Cell selection criteria	RSRP	tbd	tbd
Simulation variables	Number of UEs Backhaul properties	Number of UEs, Number of small cells	Number of UEs, Backhaul capacity
Target metric	Throughput	Throughput	Throughput / Spectral Efficiency

II.2 iJOIN evaluation scenarios

In the following, the consolidated iJOIN evaluation scenarios derived from above 3GPP scenarios are presented.

II.2.1 Outdoor deployment

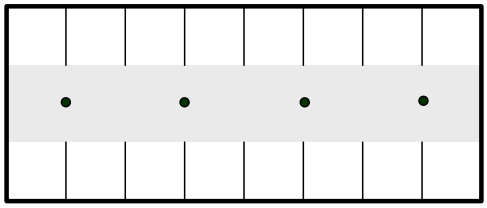
Table 6-8: iJOIN WP3 outdoor common evaluation scenario.

	Macro cell parameter	Small cell parameter
Layout	Hexagonal grid, 3 sectors per site, 19 Macro sites and 7 Macro sites can be used.	 Clusters uniformly random within macro geographical area; small cells uniformly random dropping within cluster area
System bandwidth per carrier	10MHz	10MHz
Carrier frequency	2.0GHz	2.0GHz / 3.5 GHz
Carrier number	1	1
Total BS TX power (Ptotal per carrier)	46dBm	30 dBm, Optional: 24dBm, 37dBm
Maximum UE TX power	23dBm or 24dBm	23dBm or 24dBm
Distance-dependent path loss	ITU UMa (3GPP TR36.814)	ITU Umi (3GPP TR36.814)
Shadowing (optional)	ITU UMa (3GPP TR36.819)	ITU UMi (3GPP TR36.814)
Antenna Height:	25m	10m
UE antenna Height	1.5m	
Antenna gain + connector loss	17 dBi	5 dBi (omni-directional)
Antenna gain of UE	0 dBi	
Fast fading channel between eNB and UE (optional)	ITU UMa (3GPP TR36.814); Rayleigh block fading	ITU Umi (3GPP TR36.814); Rayleigh block fading
Antenna configuration (optional)	2Tx2Rx in DL, Cross-polarized	
Number of clusters per macro cell geographical area	Depends on the common scenario	
Number of small cells per cluster	Depends on the common scenario	
Number of UEs in RRC_CONNECTED	Varying	
UE dropping	2/3 UEs randomly and uniformly dropped within the clusters, 1/3 UEs randomly and uniformly dropped throughout the macro geographical area.	
Radius for small cell dropping in a cluster	(50m) Depends on the common scenario	
Radius for UE dropping in a cluster	(70m) Depends on the common scenario	
Minimum distance (2D distance)	Small cell-small cell: 20m	
	Small cell-UE: 5m	
	Macro –small cell cluster center: 105m	

	Macro – UE : 35m	
	cluster center-cluster center: 2*Radius for small cell dropping in a cluster	
Traffic model	Full buffer	
UE speed (applies for fast and shadow fading only)	3km/h	
Cell selection criteria	iSC with best channel quality	
Backhaul model	According to table in section 5.1. (depends on common scenario and deployment option)	

II.2.2 Indoor deployment

Table 6-9: iJOIN WP3 indoor common evaluation scenario.

	Parameter
Layout	 TR36.814; 2/4 small cells per floor, 1/2 floors
System bandwidth per carrier	10MHz
Carrier frequency	2.6/3.5GHz
Carrier number	1
Total BS TX power (Ptotal per carrier)	24dBm
Distance-dependent path loss	ITU InH [referring to Table B.1.2.1-1 in TR36.814]
Penetration	0dB within the same floor;18.3dB between different floors
Shadowing	ITU InH [referring to Table A.2.1.1.5-1 in TR36.814]
Antenna Height:	6m
UE antenna Height	1.5m
Antenna gain + connector loss	5dBi
Antenna gain of UE	0 dBi
Fast fading channel between eNB and UE	ITU InH
Antenna configuration	2Tx2Rx in DL, Cross-polarized
Number of small cells per cluster	Varying
Number of UEs	Varying
UE dropping	Random
Traffic model	Full Buffer

Appendix III Categorization of Backhaul Technologies

Table 6-10 presents the outcome of the discussion which was initiated by WP3 regarding important measures for different backhaul technologies. The key parameters, like the latency, topology and throughput are the result of collaborative work between WP2-5 and input from [60]. These can be seen as guideline values for the backhaul limitations for the functional split selection. More details on this table can be found in [59].

Table 6-10: Backhaul Classification

Number	BH technology		Latency (per hop, RTT)	Throughput	Topology	Duplexing	Multiplexing Technology
1a	Millimeter wave	60GHz Unlicensed	≤5 ms	≤800 Mbit/s	PtP (LOS)	TDD	--
1b			≤200 µsec	≤1Gbps	PtP (LOS)	FDD	--
1c		70-80GHz Light licensed	≤200 µsec	≤2.5 Gbit/s	PtP (LOS)	FDD	--
2a	Microwave (28-42 GHz) Licensed		≤200 µsec	≤1Gbps	PtP (LOS)	FDD	--
2b			≤10 ms	≤1Gbps	PmP (LOS)	TDD	TDMA
3a	Sub-6 GHz Unlicensed or licensed		≤5 ms	≤500Mbps	PtP (NLoS)	TDD	--
3b			≤10 ms	≤500Mbps (shared among clients)	PmP (NLoS)	TDD	TDMA
3c			≤5 ms	≤1 Gbit/s (per client)	PmP (NLoS)	TDD	SDMA
4a	Dark Fibre		5 µs/km × 2	≤10 Gbps	PtP		--
4b	CWDM		5 µs/km × 2	≤10·N Gbps (with N≤8)	Ring		WDM
4c	Metro Optical Network		250 µs	≤1 Gbps	Mesh/Ring		Statistical Packet Multiplexing
4d	PON (Passive Optical Networks)		≤1 ms	100M – 2.5Gbps	PmP		TDM (DL)/ TDMA (UL)
5	xDSL		5-35 ms	10M – 100Mbps	PtP		--

Appendix IV CT interactions in WP3

CT 3.1	
Main functional impact	Resource Allocation
Impacted domain/resources	Backhaul/Channel
Main acting entity	RANaaS
Distributed/centralized scheme	Centralized
Specific signalling required	Yes: iSC/iTN to RANaaS for CT3.1 BH channel conditions RANaaS to central-iTN for CT3.1 BH path selection RANaaS to central-iTN for CT3.1 BH channel allocation
Operational time scale	Time scale in terms of seconds (or less)
Functional dependencies	Possible dependency with CT4.4 “Routing and Congestion Control”
Functional split constraints	Requires centralized scheduling at the RANaaS
Additional information	• CT3.1 could operate together with CT3.2 “Partly de-centralized mechanisms for joint RAN and backhaul optimization in dense small cell deployments”. CT3.2 deals with cell selection process which is un-touched during the BH routing and scheduling procedure. However, this might require coordination between CT3.2 as soon as they operate at the same time scale. • CT3.1 is not always compatible with CT3.3 “Energy-Efficient MAC/RRM at Access and Backhaul”. In particular, the discontinued transmission proposed by CT3.3 might have impact on the path selection and link scheduling process, which is mainly decided based on the channel conditions / traffic. • CT3.1 can be implemented with CT3.4 “Computational Complexity and Semi-Deterministic Scheduling”. CT3.4 performs (long-term and short term) user scheduling, whereas CT3.1 operates on top of that by assigning BH links and flows per link. These CTs do not collide; however CT3.1 can impose some additional constraints to CT3.4 for the BH availability. • CT3.1 can be implemented with CT3.5 “Cooperative RRM for Inter-Cell Interference Coordination in RANaaS”, which deals with Inter-cell RRM. These CTs do not collide; however CT3.1 can impose some additional constraints to CT3.5 for the BH availability. • CT3.1 could be implemented together with CT3.7 “Radio resource management for scalable multi-point turbo detection/In-network Processing”, since CT3.1 could be used to route traffic from users not involved in an MPTD processing. • CT3.1 is partially compatible with CT3.8 “Radio Resource Management for In-Network-Processing”, which investigates

	RRM for the uplink; however this might require coordination between the two CTs. • CT3.1 could operate with CT3.9 "Hybrid local-cloud-based user scheduling for interference control" which deals with user scheduling in downlink. These CTs do not collide; however CT3.1 can impose some additional constraints to CT3.9 for the BH availability.

CT 3.2	
Main functional impact	Connection Control
Impacted domain/resources	RAN/cell association
Main acting entity	RANaaS
Distributed/centralized scheme	Centralized
Specific signalling required	yes
Operational time scale	seconds
Functional dependencies	CTs that imply coordinated transmission and reception schemes (CT2.2/2.3/2.5) have an impact on this CT CTs where large scale scheduling is implemented are affected by this CT (CT3.4/3.7) CTs related to BH optimization 3.1 and 4.1-4.5 are affected by CT3.2
Functional split constraints	Yes, centralized connection control at the RANaaS
Additional information	• CT3.2 could operate together with CT3.1, since CT3.1 deals with small cell BH scheduling and routing while CT3.2 deals with cell association. These CTs do not collide; however, CT3.1 has to take into account the changes in cell association due to CT 3.2. • CT3.2 is fully compatible with CT3.3, since CT3.3 deals with RF transmission while CT3.2 deals with cell association. • CT3.2 could operate together with CT3.4, since CT3.1 deals with RRM. These CTs do not collide; however, long term scheduling in CT3.4 has to take into account the changes in cell association due to CT 3.2. • CT3.2 is fully compatible with CT3.5, since CT3.5 deals with short term RRM while CT3.2 deals with cell association. • CT3.2 is fully compatible with CT3.6, since CT3.6 deals modelling iJOIN network characteristics. • CT3.2 is fully compatible with CT3.7, since CT3.7 deals with short term RRM while CT3.2 deals with cell association. • CT3.2 is fully compatible with CT3.8, since CT3.8 deals with short term RRM while CT3.2 deals with cell association. • CT3.2 is fully compatible with CT3.9, since CT3.9 deals with short term RRM while CT3.2 deals with cell association.

CT 3.3	
Main functional impact	RAN RF transmission
Impacted domain/resources	RAN/RF transmission
Main acting entity	RANaaS
Distributed/centralized scheme	Centralized
Specific signalling required	yes
Operational time scale	milliseconds
Functional dependencies	CT3.2 has a tight dependency with CTs that focus on radio resource management (CT 3.4, 3.5, 3.7, 3.8, and 3.9). Cell activation and deactivation can be seen as a long term scheduling. Moreover, cooperative short term scheduling will require earlier small cell activation to enable signalling exchange. CT3.3 also affect CT3.1 since BH links can be set idle when a small cell is de-activated.
Functional split constraints	Yes, centralized connection control at the RANaaS
Additional information	• CT3.3 is fully compatible with CT3.1, since CT3.1 deals with small cell BH scheduling and routing while CT3.3 deals with RF transmission. • CT3.3 is fully compatible with CT3.2, since CT3.2 deals with cell association while CT3.3 deals with RF transmission. • CT3.3 is fully compatible with CT3.3, since CT3.2 deals with cell association while CT3.3 deals with RF transmission. • CT3.3 is compatible with CT3.4, since CT3.4 deals with RRM while CT3.3 deals with RF transmission. However, these functionalities are coupled and have to be jointly designed (coordination and signalling exchange are required) • CT3.3 is compatible with CT3.5, since CT3.5 deals with RRM/ICIC while CT3.3 deals with RF transmission. However, these functionalities are coupled and have to be jointly designed (coordination and signalling exchange are required) • CT3.2 is fully compatible with CT3.6, since CT3.6 deals modelling iJOIN network characteristics. • CT3.3 is compatible with CT3.7, since CT3.7 deals with RRM while CT3.3 deals with RF transmission. However, these functionalities are coupled and have to be jointly designed (coordination and signalling exchange are required) • CT3.3 is compatible with CT3.8, since CT3.8 deals with RRM while CT3.3 deals with RF transmission. However, these functionalities are coupled and have to be jointly designed (coordination and signalling exchange are required) • CT3.3 is compatible with CT3.9, since CT3.9 deals with RRM while CT3.3 deals with RF transmission. However, these functionalities are coupled and have to be jointly designed (coordination and signalling exchange are required)

CT 3.4	
Main functional impact	Resource Allocation
Impacted domain/resources	Backhaul/RAN/RF transmission
Main acting entity	RANaaS (and iSCs)
Distributed/centralized scheme	Centralized (partially distributed)
Specific signalling required	iSC to RANaaS: CSI (of variable granularity), pre-selection of RB allocation RANaaS to iSC: RB allocation decisions
Operational time scale	milliseconds
Functional dependencies	CT3.2, CT3.3, CT3.5, CT3.7
Functional split constraints	Scheduling entity at the RANaaS
Additional information	n/a

CT 3.5	
Main functional impact	Resource Allocation
Impacted domain/resources	RAN/Downlink radio resources
Main acting entity	RANaaS
Distributed/centralized scheme	Centralized
Specific signalling required	Yes: iSC –to-RANaaS for CT3.5 Channel State Information RANaaS-to-iSC for CT3.5 RB allocation decisions
Operational time scale	Time scale of milliseconds
Functional dependencies	No
Functional split constraints	Requires Inter-cell RRM at the RANaaS
Additional information	• CT3.5 could operate together with CT3.1 “BH Link Scheduling and QOS aware flow forwarding”, since CT3.1 deals with small cell BH scheduling and routing. These CTs do not collide; however CT3.1 can impose some additional constraints to CT3.5 for the BH availability, which might affect the Inter-cell RRM. • CT3.5 could operate together with CT3.2 “Partly de-centralized mechanisms for joint RAN and backhaul optimization in dense small cell deployments”. CT3.2 deals with cell selection process which is un-touched during the proposed ICIC. • CT3.5 is not always compatible with CT3.3 “Energy-Efficient MAC/RRM at Access and Backhaul”, since it deals also with RRM for small cells from different perspective (having different objective). • CT3.5 cannot be implemented with CT3.4 “Computational Complexity and Semi-Deterministic Scheduling”. CT3.4 performs (long-term and short term) user scheduling and this might collide with CT3.5, which provides a multi-cell user

	scheduling solution in downlink. • CT3.5 could not be implemented together with CT3.7 "Radio resource management for scalable multi-point turbo detection/In-network Processing", since CT3.5 performs RRM in a systematic manner for all the users in a cluster of small cells (needs discussion). • CT3.5 is compatible with CT3.8 "Radio Resource Management for In-Network-Processing", which investigates RRM for the uplink. • CT3.5 could not operate with CT3.9 "Hybrid local-cloud-based user scheduling for interference control" which deals also with the user scheduling in downlink as CT3.5.

CT 3.7	
Main functional impact	Resource Allocation (large scale scheduling)
Impacted domain/resources	RAN/Uplink radio resources
Main acting entity	RANaaS (and iSCs)
Distributed/centralized scheme	Centralised scheme
Specific signalling required	iSC to RANaaS for CT3.7 activation request RANaaS to iSC for CT3.7 information request iSC to RANaaS for CT3.7 information response RANaaS to iSC for CT3.7 activation response RANaaS to iSC for CT3.7 parameters (resource allocation) iSC to RANaaS for CT3.7 deactivation request (tentative) RANaaS to iSC for CT3.7 deactivation confirmation (tentative)
Operational time scale	General framework update could be done every second (less is better through)
Functional dependencies	CT2.2
Functional split constraints	No. However, if CT2.2 processing is done in RANaaS, then functional split at PHY layer after iFFT is preferred
Additional information	• CT3.7 could be implemented together with CT3.1 "Backhaul Link Scheduling and QoS-aware Flow Forwarding", since CT3.1 deals with backhaul routing to the core network essentially. CT3.1 would be used to route traffic from users not involved in an MPTD processing (no side effect so far) • CT3.7 could be implemented with CT3.2 "Partly decentralized mechanisms for joint RAN and backhaul optimization in dense small cell deployments", since CT3.2 deals with cell (re) selection mechanism. CT3.7 assumes the selection is done, while CT3.2 will act on the selection before CT3.7 has to be applied (no side effect so far). • CT3.7 may not be compatible with CT3.3 "Energy-Efficient MAC/RRM at Access and Backhaul" which deals with discontinuous transmission of iSCs in the downlink. CT3.7 requires that the identified iSCs stay up (discussion is needed).

	• CT3.7 may not be implemented with CT3.4 "Computational Complexity and Semi-Deterministic Scheduling", which deals with RRM in a centralised way: long term scheduling done by the RANaaS, while short term scheduling operated at each iSC. CT3.7 is also a centralised RRM CT and is a "concurrent" of CT 3.4. Ideally if CT3.4 only deals with UEs not involved in MPTD, while CT3.7 operates on those specific UEs, then CT3.7 could be implemented together. (discussion is needed) • CT3.7 could be implemented with CT3.5 "Cooperative RRM for Inter-Cell Interference Coordination in RANaaS" which deals with downlink RRM (no side effect so far). • CT3.7 could be implemented with CT3.6 "Utilization and Energy Efficiency" which evaluates those metrics with the iJOIN context (no side effect so far). • CT3.7 may not be compatible with CT3.8 "Radio Resource Management for In-Network-Processing", which is a concurrent uplink RRM method (discussion is needed). • CT3.7 could be implemented with CT3.9 "Hybrid local-cloud-based user scheduling for interference control" which deals with scheduling in the downlink, while CT3.7 operates in the uplink (no side effect so far).

CT 3.8	
Main functional impact	Resource Allocation
Impacted domain/resources	RAN/Uplink radio resources
Main acting entity	RANaaS, iveC
Distributed/centralized scheme	centralized
Specific signalling required	yes
Operational time scale	typical scheduling time scale
Functional dependencies	CT2.1
Functional split constraints	split within PHY between detection and decoding or between PHY and MAC
Additional information	• CT3.8 may be compatible with CT3.1, because the jointly detected user data symbols or bits need to be forwarded to the RANaaS over the backhaul network. Nevertheless side effects need to be investigated and in general, coordination is required • CT3.8 can be combined with CT3.2, but in addition to a primary cell association (for control channels), also an additional assignment of jointly detecting small cells is performed by CT3.8, which needs to be coordinated • CT3.8 can be combined with CT3.3, since DTX can be considered as a RRM technique • CT3.8 is not compatible with CT3.4, since CT3.8 assumes centralized RRM

	• CT3.8 can be combined with CT3.5, since CT3.8 considers the uplink only, while CT3.5 considers only downlink • CT3.8 is not compatible with CT3.7 since it relies on CT2.1, which is an alternative to CT2.2 (on which CT3.7 relies) • CT3.8 is compatible with CT3.9, since it operates on uplink only, while CT3.9 considers downlink only

CT 3.9	
Main functional impact	Resource Allocation
Impacted domain/resources	Backhaul/RAN/RF transmission
Main acting entity	iSCs (potential extension with RANaaS)
Distributed/centralized scheme	Distributed (potential extension to partially centralized)
Specific signalling required	iSC-iSC, iSC-RANaaS: CSI short terms (when possible), specific signalling (when possible), long term CSI
Operational time scale	scheduling time, specific signalling
Functional dependencies	CT3.1, CT3.2, CT3.3,
Functional split constraints	Scheduling entity at the iSCs
Additional information	• CT3.1 (Backhaul link scheduling and QoS-aware flow forwarding) : • CT3.2 (Partly de-centralized mechanisms for joint RAN and backhaul optimization in dense small cell deployment) : This CT deals with cell selection and can operate with CT3.5 as it works on a different level. • CT3.3 (Energy-Efficient MAC/RRM at Access and Backhaul) optimizes the activation and deactivation of cells and can operate with CT3.5 as it works on a different level. • CT3.4 (Computation complexity and semi-deterministic scheduling). It is impossible to apply both CTs because both offer alternative solutions for different settings and are operating on the same resources. • CT3.5 (Cooperative RRM for Inter-Cell Interference Coordination in RANaaS) It is impossible because both offer alternative solutions for different settings and are operating on the same resources. • CT3.7 (Radio resource management for scalable multi-point turbo detection/In-network Processing). This CT operates on the uplink and is compatible with CT3.9 which operates on the downlink. • CT3.8 (Radio Resource Management for In-Network-Processing) This CT operates on the uplink and is compatible with CT3.9 which operates on the downlink.

References

[1] S. Parkvall, et al., “Heterogeneous network deployments in LTE”, Ericsson Review, vol. 2, 2011.

[2] 3GPP, “TR 25.814 v7.1.0; Physical Layer Aspects for Evolved UTRA”, Sep. 2006.

[3] 3GPP, “TR 36.932 V12.0.0; Scenarios and Requirements for Small Cell Enhancements for E-UTRA and E-UTRAN”, Dec. 2012

[4] P. Rost, et al., “Cloud Technologies for Flexible 5G Radio Access Networks”,, IEEE Communications Magazine, May 2014.

[5] INFSO-ICT-317941 iJOIN, D3.1, “Final report on MAC/RRM state-of-the-art, Requirements, scenarios and interfaces in the iJOIN architecture“, Nov. 2013. [available online] http://www.ict-ijoin.eu/wp-content/uploads/2014/01/D3.1.pdf

[6] D. Pisinger, “Algorithms for knapsack problems”, Ph.D. dissertation, University of Copenhagen, 1995.

[7] A. De Domenico, V. Savin, and D. Kténas, “Cell Selection for Joint Optimization of the Radio Access and Backhaul in Heterogeneous Cellular Networks”, submitted to IEEE Transactions on Wireless Communications.

[8] R. S Sutton and A. G Barto, “Reinforcement learning: An introduction, vol. 1”, Cambridge Univ. Press, 1998.

[9] D.V. Djonin and V. Krishnamurthy, “MIMO transmission control in fading channels-a constrained markov decision process formulation with monotone randomized policies”, IEEE Transactions on Signal Processing, vol. 55, no. 10, pp. 5069–5083, 2007.

[10] M. Bkassiny, Yang Li, and S.K. Jayaweera, “A survey on machine-learning techniques in cognitive radios”,, IEEE Communications Surveys Tutorials, vol. 15, no. 3, pp. 1136–1159, 2013.

[11] A. Galindo-Serrano and L. Giupponi, “Distributed q-learning for aggregated interference control in cognitive radio networks”, IEEE Transactions on Vehicular Technology,, vol. 59, no. 4, pp. 1823–1834, 2010.

[12] M.A. Wiering and E.D. de Jong, “Computing optimal stationary policies for multi-objective Markov decision processes”, in IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning 2007, pp. 158–165, 2007.

[13] L. Saker, S.-E. Elayoubi, R. Combes, and T. Chahed, “Optimal control of wake up mechanisms of femtocells in heterogeneous networks”, IEEE Journal on Selected Areas in Communications, vol. 30, no. 3, pp. 664–672, 2012.

[14] INFSO-ICT-317941 iJOIN, D2.1, “State-of-the-art of and promising candidates for PHY layer approaches on access and backhaul network“, Nov. 2013. [available online] http://www.ict-ijoin.eu/wp-content/uploads/2014/01/D2.1.pdf

[15] INFSO-ICT-317941 iJOIN, D5.1, “Revised definition of requirements and preliminary definition of the iJOIN architecture“, Nov. 2013. [available online] http://www.ict-ijoin.eu/wp-content/uploads/2014/01/D5.1.pdf

[16] 3GPP TS 36.420, “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 general aspects and principles (Release10)”, V10.2.0, Sept. 2011.

[17] 3GPP TS 36.421, “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 layer 1 (Release10)”, V10.0.1, Mar. 2011.

[18] 3GPP TS 36.422, “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 signalling transport (Release10)”, V10.1.0, Jun. 2011.

[19] 3GPP TS 36.423, “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 Application Protocol (X2AP) (Release10)”, V10.7.0, Sept. 2013.

[20] 3GPP TS 36.424, “Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 data transport (Release10)”, V10.1.0, Jun. 2011.

[21] 3GPP TR 36.872, “Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects (Release12)”, V12.1.0, Dec. 2012.

[22] ITU-R M2135, “Report ITU-R M2135, Guidelines for evaluation of radio interface technologies for IMT-Advanced”, 2008.

[23] 3GPP TR 36.814, “Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects (Release9)”, V9.0.0, Mar. 2010.

[24] S. M. Kay, “Fundamentals of Signal Processing: Estimation Theory”, Prentice Hall PTR, 1st Edition, 1993.

[25] Hui Zhou, Pingyi Fan, and Jie Li, "Global Proportional Fair Scheduling for Networks With Multiple Base Stations", IEEE Transactions on Vehicular Technology, vol.60, no.4, pp.1867-1879, May 2011.

[26] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan, “The Evolution of Cellular Backhaul Technologies: Current Issues and Future Trends”, IEEE Communications Surveys & Tutorials, vol.13, no.1, pp.97-113, First Quarter 2011.

[27] Alcatel-Lucent, “Leveraging VDSL2 for mobile backhaul”, white paper, 2010.

[28] “Small Cell Backhaul Requirements,” white paper, Next Generation Mobile Networks (NGMN) Alliance, June. 2012.

[29] S. Chia, M. Gasparroni, and P. Brick, “The next challenge for cellular networks: Backhaul”, IEEE Microwave Magazine, vol.10, no.5, pp.54-66, August 2009.

[30] L. Yinggang, "E-band radios for LTE/LTE-Advanced mobile backhaul", 2010 Workshop on Integrated Nonlinear Microwave and Millimeter-Wave Circuits (INMMIC), pp.84, April 2010.

[31] D. Wübben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, “Benefits and Impact of Cloud Computing On 5G Signal Processing”, submitted to IEEE Signal Processing Magazine.

[32] 3GPP, “TR 36.300 V11.4.0; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description”, Dec. 2012.

[33] M. Fischetti, J. J. S. Gonzlez, and P. Toth, “A branch-and-cut algorithm for the symmetric generalized traveling salesman problem”, Operations Research, vol. 45, no. 3, pp. pp. 378–394, 1997.

[34] K. Kim; Y. Han and S-L. Kim, "Joint subcarrier and power allocation in uplink OFDMA systems,” IEEE Communications Letters, vol.9, no.6, pp. 526-528, Jun 2005.

[35] Li Hongjia,X. Xiaodong, H. Dan, Q. Xiqiang, T. Xiaofeng and Z. Ping, "Graph Method Based Clustering Strategy for Femtocell Interference Management and Spectrum Efficiency Improvement", 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), pp.1-5, Sept. 2010.

[36] R. Y. Chang, T. Zhifeng, J. Zhang and C. C-J Kuo, "Multicell OFDMA Downlink Resource Allocation Using a Graphic Framework," IEEE Transactions on Vehicular Technology, vol.58, no.7, pp.3494-3507, Sept. 2009.

[37] M. Sharif and B. Hassibi, “On the capacity of MIMO broadcast channels with partial side information”, IEEE Transactions on Information Theory, Feb. 2005.

[38] P. de Kerret, S. Lasaulce, D. Gesbert, and U. Salim, “Best-response team power control for the interference channel with local CSI”, submitted to ICC 2015.

[39] R. Radner, “Team decision problems”, the Annals of Mathematical Statistics, 1962.

[40] J. Nash, “Non-cooperative games”, Annals of Mathematics, 1951.

[41] M. Ebrahimi, M. A. Maddah-Ali, and A. K. Khandani, “Throughput scaling laws for wireless networks with fading channels”, IEEE Transactions on Information Theory, vol.53, no. 11, pp.4250-4254, Nov. 2007.

[42] D. Wübben, H. Paul, B.-S. Shin, G. Xu, and A. Dekorsy, "Distributed Consensus-based Estimation for Small-Cell Cooperative Networks," European Conference on Networks and Communications (EuCNC), Bologna, Italy, June 2014.

[43] INFSO-ICT-317941 iJOIN, D2.2, "Definition of PHY layer approaches that are applicable to RANaaS and a holistic design of backhaul and access network ", November 2014.

[44] 3GPP, "TS 36.321 V10.7.0; Medium Access Control (MAC) protocol specification", Dec. 2012

[45] 3GPP, "TS 36.322 V10.0.0; Radio Link Control (RLC) protocol specification", Dec. 2010

[46] 3GPP, "TS 36.323 V10.2.0; Packet Data Convergence Protocol (PDCP) specification", Dec. 2012

[47] 3GPP, "TS 36.331 V10.8.0; Radio Resource Control (RRC); Protocol specification", Dec. 2012

[48] S. Feng and E. Seidel, "Self-organizing networks (SON) in 3GPP long term evolution," Nomor Research GmbH, White Paper, 2008.

[49] M. Nohrborg, "Self-Organizing Networks," [available online] http://www.3gpp.org/technologies/keywords-acronyms/105-son, last access December 2013.

[50] H. Guan, T. Kolding, and P. Merz, "Discovery of Cloud-RAN," in Cloud-RAN Workshop, April 2010.

[51] Wind River, "Virtualization Performance Delivered", [available online] http://www.windriver.com/announces/open_virtualization/wind-river-ov_benchmarks.pdf, last access April 2014.

[52] ETSI NFV ISG, "Network Function Virtualization – Update White Paper", [available online] http://portal.etsi.org/NFV/NFV_White_Paper2.pdf, October 2013

[53] D. Wübben, H. Paul, B.-S. Shin, G. Xu, and A. Dekorsy, "Distributed Consensus-Based Estimation for Small Cell Cooperative Networks" Globecom 2014 Workshop – Broadband Wireless Access, Austin, USA, Dec. 2014 (accepted)

[54] D. Naddef and G. Rinaldi, "Branch-and-cut algorithms for the capacitated VRP", In book: *The Vehicle Routing Problem*, Society of Industrial and Applied Mathematics, 2001, pp.29-51.

[55] L. Tassiulas and A. Ephremides, "Stability Properties of Constrained Queuing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks", IEEE Tran. on Automatic Control, 37(12):1936–1948, Dec. 1992.

[56] M.R. Akdeniz et al., "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *Selected Areas in Communications, IEEE Journal on* , vol.32, no.6, pp.1164,1179, June 2014 doi: 10.1109/JSAC.2014.2328154.

[57] B. Tianyang and R.W. Heath, "Coverage analysis for millimeter wave cellular networks with blockage effects," *Global Conference on Signal and Information Processing, 2013 IEEE* , vol., no., pp.727,730, 3-5 Dec. 2013.

[58] A. De Domenico, E. Calvanese Strinati, and A. Capone, "Enabling green cellular networks: A survey and outlook," Elsevier Computer Communications, vol. 37, pp. 5 – 24, 2014.

[59] INFSO-ICT-317941 iJOIN, D4.2, "Network-layer algorithms and network operation and management: candidate technologies specification ", November 2014.

[60] Small Cell Forum, "Backhaul Technologies for Small Cells - Use Cases, Requirements and Solutions", February 2013, http://ytd2525.files.wordpress.com/2013/03/049-backhaul-technologies-small-cells.pdf

[61] NTT DOCOMO, NEC, "R3-142122; CHANGE REQUEST to TS 36.300; Introduction of Dual Connectivity", Oct. 2014

[62] Small Cell Forum, "LTE eNB L1 API definition", February 2014, Document 082.04.01, Release 4.

[63] INFSO-ICT-317941 iJOIN, D5.2, "Final Definition of iJOIN Requirements and Scenarios", November 2014.P. Rost, S. Talarico, M. C. Valenti, *The Complexity-Rate Tradeoff of Centralized Radio Access Networks*, submitted to IEEE Transactions on Wireless Communications

[64] ITU-R, "Rep. ITU-R M.2135; Guidelines for evaluation of radio interface technologies for IMT-Advanced", 2008

[65] 3GPP, "TS 36.213 V10.4.0; Physical layer procedures", Dec. 2011

[66] The Green Grid, " PUE[tm]: A comprehensive Examination of the Metric", 2012, [available online] http://www.thegreengrid.org/~/media/WhitePapers/WP49-PUE%20A%20Comprehensive%20Examination%20of%20the%20Metric_v6.pdf?lang=en

[67] ETSI ISG NFV - NFV Requirements – Liason Statement directed to NFV OpenStack Team, September 2014, [available online] https://wiki.openstack.org/w/images/c/c7/NFV%2814%29000154r2_NFV_LS_to_OpenStack.pdf .

[68] P. Rost and A. Prasad, "Opportunistic Hybrid ARQ – Enabler of Cloud-RAN over Non-Ideal Backhaul," IEEE Wireless Communications Letters, June 2014.

[69] E. Pateromichelakis, M. Shariat, A. Quddus and Tafazolli, R., "Graph-Based Multicell Scheduling in OFDMA-Based Small Cell Networks," *Access, IEEE* , vol.2, no., pp.897,908, 2014.

[70] P. Rost and A. Prasad, "Opportunistic Hybrid ARQ—Enabler of Centralized-RAN Over Nonideal Backhaul," *Wireless Communications Letters, IEEE* , vol.3, no.5, pp.481,484, Oct. 2014.

[71] A. Prasad and A. Maeder, "Backhaul-aware Energy Efficient Heterogeneous Networks with Dual Connectivity", Springer Telecommunication Systems (to appear).

[72] A. Maeder, et al., "Towards a flexible functional split for cloud-RAN networks," *Networks and Communications (EuCNC), 2014 European Conference on* , vol., no., pp.1,5, 23-26 June 2014.

[73] A. Prasad and A. Maeder, "Energy Saving Enhancement for LTE-Advanced Heterogeneous Networks with Dual Connectivity*", IEEE VTC Fall*, 2014.

[74] M. Valenti, S. Talarico and P. Rost, "The Role of Computational Outage in Dense Cloud-Based Centralized Radio Access Network", *IEEE Globecom 2014* (to appear).

[75] R. Fritzsche, P. Rost and G. P. Fettweis, "Robust Proportional Fair Scheduling with Imperfect CSI and Fixed Outage Probability", *IEEE PIMRC 2014* (to appear)

