iJOIN
INFSO-ICT-317941

INFSO-ICT-317941 iJOIN

Deliverable D5.3

Final definition of iJOIN architecture

Editor:	Antonio De Domenico (CEA); Peter Rost (NEC)
Deliverable nature:	Public
Suggested readers:	iJOIN General Assembly
Due date:	April 30th , 2015
Delivery date:	April 30th , 2015
Version:	1.0
Total number of pages:	78
Reviewed by:	GA members
Keywords:	RAN, Backhaul, Cloud Computing, RANaaS, iJOIN Common Scenarios, iJOIN Candidate Technology, iJOIN Architecture, LTE
Resources consumed	8.81 PM

Abstract

This document provides a conclusive holistic description and evaluation of the iJOIN concepts by underlying the advantages of the iJOIN framework with respect to the existing technologies. In particular iJOIN aims at finding innovative solutions to satisfy the high data rate requirements of future wireless networks through dense small cell deployments with heterogeneous backhaul. iJOIN has introduced the concept of "Radio Access Network as a Service" (RANaaS), where RAN functionalities are partially or fully executed in a cloud platform to improve the system performance in terms of throughput, energy efficiency, cost efficiency, and utilization efficiency.

At first, this deliverable recalls the baseline Long Term Evolution (LTE) architecture, and presents the finalised iJOIN logical/functional/physical architectures developed under the joint collaboration of the different work packages. Then, we focus on the final evaluation of iJOIN key enablers, the flexible functional split and the design of joint operation at the radio access and backhaul networks. These concepts are introduced and investigated to demonstrate their potential gains with respect to the baseline technologies.

Finally, to evaluate the iJOIN framework performance with respect to the key objectives, this deliverable provides a project-wide analysis of the different candidate technologies proposed in iJOIN on four representative scenarios.

List of authors

Company	Author
CEA	Antonio De Domenico (antonio.de-domenico@cea.fra)
HP	Marco Di Girolamo (marco.digirolamo@hp.com)
	Marco Consonni (marco_consonni@hp.com)
IMC	Umer Salim (umer.salim@intel.com)
IMDEA	Jorge Ortín (jortin@unizar.es)
	Albert Banchs (albert.banchs@imdea.org)
	Pablo Caballero (pablo.caballero@imdea.org)
NEC	Andreas Maeder (andreas.maeder@neclab.eu)
	Peter Rost (peter.rost@neclab.eu)
SCBB	Massinissa Lalam (massinissa.lalam@sagemcom.com)
TI	Dario Sabella (dario.sabella@telecomitalia.it)
	Marco Caretti (marco.caretti@telecomitalia.it)
TID	Ignacio Berberana (ibfm@tid.es)
TUD	Jens Bartelt (jens.bartelt@ifn.et.tu-dresden.de)
UC3M	Carlos Jesús Bernardos Cano (cjbc@it.uc3m.es)
UNIS	Atta Quddus (a.quddus@surrey.ac.uk)
	Yingli Sheng (y.sheng@surrey.ac.uk)
UoB	Dirk Wübben (wuebben@ant.uni-bremen.de) Henning Paul (paul@ant.uni-bremen.de)

History

Modified by	Date	Version	Comments
Antonio De Domenico	April 30th, 2015	1.0	Final version of D5.3

Table of Contents

List of Figures

List of tables

Abbreviations

3GPP	3rd Generation Partnership Project
A/D	Analog-to-digital
ARQ	Automatic Repeat Request
BH	Backhaul
BS	Base Station
CAPEX	Capital Expenditures
CBF	Constant bit rate flows
CDF	Cumulative Distribution Function
CoMP	Cooperative Multi-Point
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
C-RAN	Centralized RAN
CSI	Channel State Information
CT	Candidate Technology
DL	Downlink
DTX	Cell discontinuous transmission
ECN	Explicit Congestion Notifications
EE	Energy Efficiency
EM	Element Management
eNB	Evolved Node B
EPC	Evolved Packet Core
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FDD	Frequency Division Duplexing
FEC	Forward Error Correction
FFT	Fast Fourier Transform
GPP	General Purpose Processor
HARQ	Hybrid Automatic Repeat Request
IaaS	Infrastructure as a Service
ICIC	Inter-Cell Interference Coordination
iLGW	iJOIN Local Gateway
iNC	iJOIN Network Controller
IP	Internet Protocol
IQ	In-phase and quadrature
iRPU	iJOIN virtual RAN Processing Units
iSC	iJOIN Small Cell
IT	Information Technology
iTN	iJOIN Transport Node
iveC	iJOIN virtual eNodeB Controller
JNCC	Joint network-channel coding approach
KVM	Kernel-based Virtual Machine
LDPC	Low-density parity check
LOS	Line-of-Sight
LTE	Long Term Evolution
MAC	Medium Access Control
MANO	Management and Orchestration
MCS	Modulation and Coding Scheme
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MPLS	Multi-Protocol Label Switching
MPTD	Multi-Point Turbo Detection
MUD	Multi-User Detection
NFV	Network Function Virtualization
NFVI	NFV Infrastructure
OAM	Operations, Administration, and Management
OPEX	Operational Expenditures
OS	Operating System

OSS	Operations Support System
PDCP	Packet Data Convergence Protocol
PDU	Packet Data Unit
PHY	Physical Layer
QoS	Quality of Service
OSS	Operation Support System
RAM	Random Access Memory
RAN	Radio Access Network
RANaaS	Radio Access Network as a Service
RAT	Radio Access Technology
RF	Radio Frequency
RLC	Radio Link Control
RoF	Radio over Fibre
RRC	Radio Resource Control
RRM	Radio Resource Management
RRH	Remote Radio Head
RTT	Round-Trip Time
SDN	Software Defined Networking
S-GW	Serving Gateway
SNR	Signal-to-Noise Ratio
TCO	Total Cost of Ownership
TCP	Transport Control Protocol
UE	User Equipment
UK	United Kingdom
UL	Uplink
UM	Unacknowledged Mode
VBF	Variable bit rate flows
veNB	virtual evolved Node B
VM	Virtual Machine
VNF	Virtual Network Functions
VNFC	VNF components
WP	Work Package

1 Executive Summary

This document provides a conclusive description of the work in Work Package 5. At first, it gives the final definition of the iJOIN architecture and a summary of reference scenarios and system requirements considered in iJOIN. In particular, D5.3 has the goal to define the final iJOIN system concept and to provide a final evaluation of the proposed technologies with respect to the key objectives.

WP5 provides a description of the iJOIN system integrating the bits and pieces that were developed in WP2, WP3, and WP4, leveraging the proof of concept of WP6 and complementing the CTs with the necessary architecture. It further shows how iJOIN's key concepts are applied to mobile networks and how technologies described by WP2, WP3, and WP4 are integrated into these concepts. More specifically, this deliverable elaborates on the system design and a quantitative project-wide performance evaluation with respect to area throughput, energy-efficiency, cost-efficiency, and utilisation-efficiency.

The first part of this deliverable, Section 3, provides a revised set of terms and definitions considered as a reference in the iJOIN project. Section 4 describes the iJOIN system, key enablers, key concepts, and iJOIN design criteria. The iJOIN system design criteria are important to understand how iJOIN technologies are applied and combined to achieve the project's objectives. The design criteria focus on the individual trade-offs that are present in a mobile network and how these trade-offs can be exploited.

Section 5 gives a detailed overview of the final iJOIN architecture. The section summarises the iJOIN logical architecture, its logical elements, and how these elements interact. Then, the RANaaS cloud architecture is described which defines how the RANaaS system is implemented on cloud-computing hardware. Furthermore, we describe the iJOIN functional architecture which defines the interactions of individual iJOIN technologies within and across the work packages. Finally, the iJOIN physical architecture is described which is an application of the iJOIN logical architecture to the different common scenarios.

Section 6 details the functional split concepts. It provides a comprehensive overview of the concepts by describing the main implementation requirements, i.e., mainly computational and timing constraints, and major gains such as aggregation gains related to different functional split solutions as well as computational diversity gains.

Section 7 provides a description of possible gains and requirements of joint RAN/BH design and operation. In particular, this section elaborates on a joint RAN/BH architecture with relevant interfaces required to support iJOIN technologies. Furthermore, novel technologies such as joint RAN/BH network energy optimisation, joint RAN/BH channel coding and congestion control are described.

Section 8 provides a quantitative project-wide system evaluation with a description of the evaluation methodology and the subsequent presentation of system-wide results based on four representative scenarios. Furthermore, Section 8 elaborates on the cost-efficiency and utilisation-efficiency of the iJOIN system concept.

2 Introduction and key contributions

2.1 Motivation and Background

The iJOIN project investigates solutions to support ultra-dense small cell deployments in future cellular wireless networks. To limit the impact of the inter-cell interference and to deal with heterogeneous small-cell backhaul, iJOIN has proposed to incorporate flexibly centralised Radio Access Network (RAN) functionality into a legacy LTE architecture. According to the new RAN as a Service (RANaaS) paradigm, the centralisation of the RAN functions depends on the actual system requirements as well as network constraints. In order to implement the iJOIN vision, we have designed joint RAN/Backhaul (BH) operations [8][11]. Moreover, we have introduced a Software-Defined Networking (SDN) architecture to adapt the transport network to the momentary functional split and actual service characteristics [14].

The previous deliverable D5.1 [1] has described the baseline LTE architecture as well as state of the art, and it provided the basic knowledge for understanding the iJOIN concepts and technologies. Then, D5.2 [15] has described in detail the requirements and the constraints for the smooth implementation of the two iJOIN enablers, functional split and joint RAN/BH operation. In addition, D5.2 detailed the most relevant use cases for future generation of cellular networks. Our preliminary results have shown that the iJOIN paradigm can improve the system performance in terms of computational burden, energy efficiency, and user performance [15]. Moreover, we have described in D5.2 [15] a harmonized evaluation methodology to provide a consistent and holistic evaluation of our proposals with respect to the four iJOIN quantitative objectives: area throughput, energy efficiency, utilization efficiency, and cost efficiency.

Based on this evaluation methodology and starting from the outcome of previous work, the overall system concept is described in this deliverable. Moreover, the efficiency of functional split and joint RAN/BH optimization concepts are demonstrated by evaluating the combination of different technologies and related system-wide performance gains obtained compared to baseline technologies. This harmonization is the main task of Work Package 5 (WP 5), which takes as input the solutions developed in WP 2, WP 3, and WP 4 where novel solutions for physical layer, medium access and radio resource control, and for the transport network are derived, respectively.

2.2 Key Contributions

In the following, we provide a list of the key contributions which are described in this document:

- A detailed description of system design criteria which takes into account the inherent trade-offs of iJOIN's key objectives as well as the different operating regimes of technologies investigated in iJOIN's work packages,
- An update of the iJOIN logical, functional, and physical architectures; these architecture have been partially presented in IEEE Communications Magazine [45], and IEEE Open Access [19],
- A description of the iJOIN RANaaS architecture which defines how centralized components interact in a cloud-computing environment and which is in line with the ETSI NFV architecture,
- A detailed analysis and description of preferred functional splits, their expected gains and their requirements on the radio access and backhaul network,
- A detailed analysis and description of the technologies that optimise RAN and BH jointly; a particular emphasis has been paid to the impact of RANaaS functional splits on the backhaul network [18]
- A detailed cost-efficiency evaluation of capital expenditures in distributed RAN (conventional implementation), C-RAN, and iJOIN's RANaaS concept which has been partially presented in IEEE Journal on Selected Areas on Communications [37],
- A detailed analysis of the computational requirements of a 3GPP LTE implementation based on iJOIN's RANaaS concept which has been submitted to IEEE Transactions on Wireless Communications [35] (as minor revision),
- The final iJOIN system evaluation which takes into account performance results from all work package with project-wide defined scenarios in order to quantify the performance with respect to iJOIN's key objectives.

3 Definitions

This section presents the concepts and definitions used within iJOIN to guarantee a common understanding to all partners and external readers about the terminology used across the document. A general definition of the relevant parts of mobile and backhaul radio networks is shown in Figure 3-1.

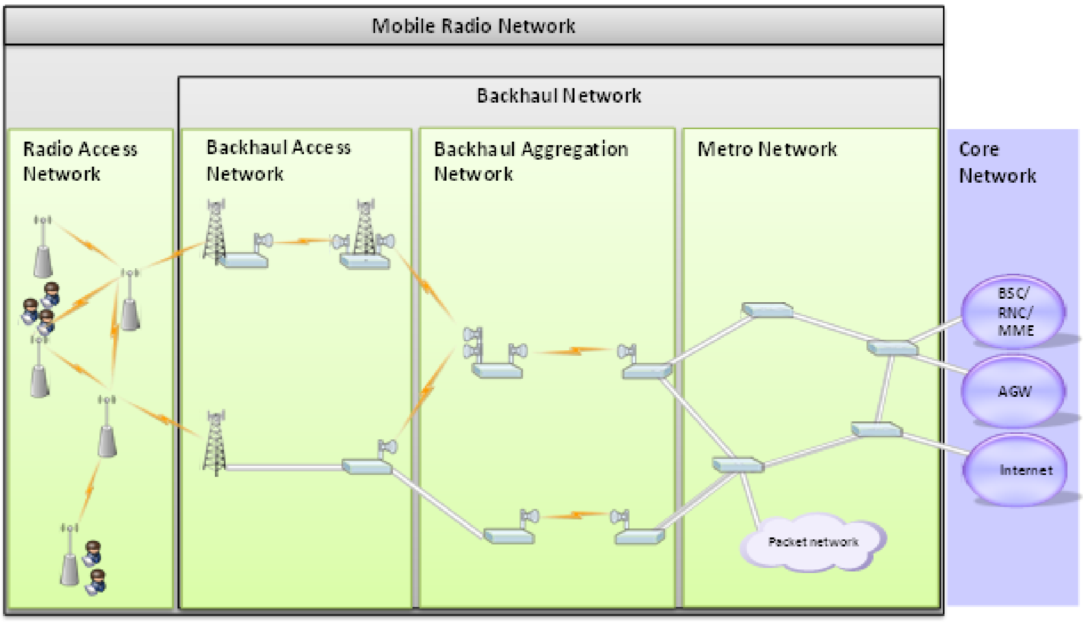

Figure 3-1: Mobile Radio Network Definition

3.1 General Nomenclature

Radio Access (RA): Wireless radio frequency (RF) link between the User Equipment (UE) and the Radio Access Network (RAN).

Radio Access Network (RAN): Network elements and functions required to support the Radio Access operation (E-UTRAN).

Small Cell (SC): Low power base station with intelligence, part of the Radio Access Network. A Small Cell

- is an operator-controlled equipment;
- supports the same functionalities as an evolved Node B (eNB);
- can be deployed indoors or outdoors;
- can be within or outside the coverage of a macro-cell.

Radio Remote Head (RRH): Radio frequency processing unit without intelligence (e.g. optical to radio conversion), i.e. Radio-over-Fibre.

Backhaul (BH): Links connecting the Radio Access Network (E-UTRAN) and the Core Network (EPC).

Backhaul Network: Network elements and functions required to support the backhaul operation.

Fronthaul (FH): Link within the Radio Access Network allowing for a distributed implementation of the RF layer of a base station, e.g. optical link between baseband processing units and RRHs. As a convention in iJOIN, all links within the Radio Access Network allowing for a distributed implementation of the upper OSI layer(s) (L1/L2/L3) will be referred to as backhaul (see Figure 3-2).

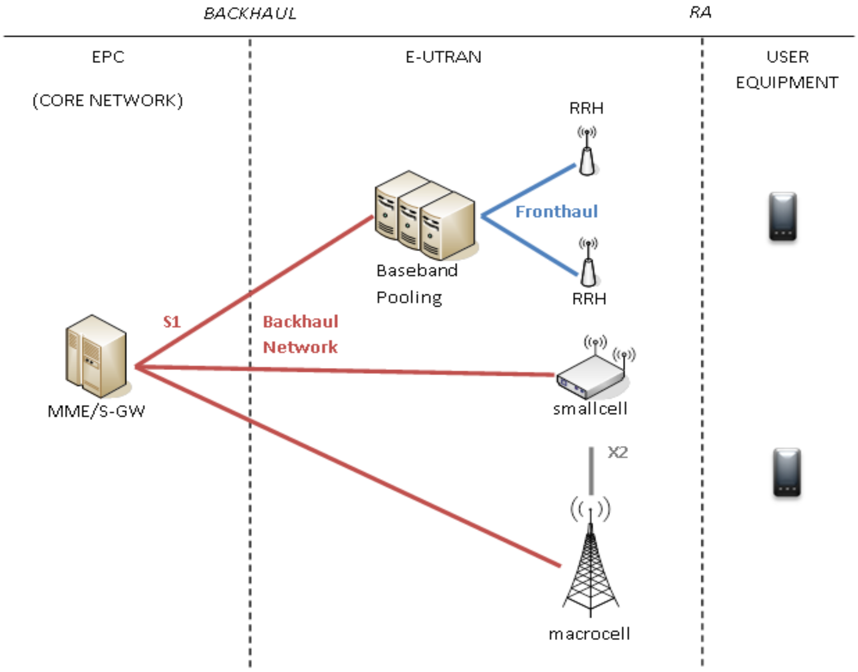

Figure 3-2: "Generic" Mobile Network Architecture

3.2 iJOIN Specific Nomenclature

RAN as a Service (RANaaS): One key concept introduced by iJOIN. It indicates a set of computing and storage infrastructure resources (typically a cloud-computing Infrastructure as a Service (IaaS) platform, but potentially any industry standard, general purpose computing infrastructure) where, according to the functional split concept investigated by iJOIN, part of the RAN protocol functionality executed.

RANaaS platform: Technology baseline implementing one or more specific instances of RANaaS. For example, in case of IaaS based RANaaS, the RANaaS platform encompasses the physical infrastructure resources (servers, storage, and their connectivity) and the corresponding cloud management software.

RANaaS instance: A specific implementation of a RANaaS platform, serving a set of iJOIN small cells and executing the upper domain of a virtual eNB.

RANaaS Point of Presence (RANaaS-PoP or iPOP): Physical or virtual location where one or more RANaaS instances are deployed and executed. Examples of RANaaS-POP locations might include fully owned enterprise data-centres, co-located data-centres, and virtual private clouds. Different RANaaS instances can potentially be placed in a common RANaaS-PoP, assuming there are mechanisms ensuring a full separation among them.

iJOIN Small Cell (iSC): Logical entity introduced by iJOIN. Low power flexible radio access point implementing fully or partially the RAN protocol stack, while upper layers may be handled by the RANaaS instance. Apart from the functional split concept introduced by iJOIN, an iSC has all the same properties of a standard small cell. An iSC is connected to a RANaaS instance through the logical J1 interface, and to another iSC through the logical J2 interface.

virtual eNB (veNB): Logical entity enclosing the set of functions and interfaces which, in the iJOIN architecture, correspond to the implementation of an eNodeB according to 3GPP specifications. A veNB is composed by a RANaaS instance and one or more iSCs. Functions and interfaces are not necessarily executed or placed in the same physical or logical network entity.

iJOIN virtual RAN Processing Unit (iRPU): Logical sub-entity of a virtual eNB, located in the RANaaS instance and executing the centralized RAN functionality of a virtual eNB. The interworking of iRPUs and iSCs therefore constitutes the realisation of a specific functional split.

iJOIN veNB Controller (iveC): Logical sub-entity of a veNB, located in the RANaaS platform, responsible for functional distribution across the veNB, consistent execution of the distributed functionalities, management and configuration of the different veNB components.

iJOIN Network Controller (iNC): Logical entity in charge of controlling the joint RAN/BH operation. To minimise the impacts for the operator in terms of deployment cost and complexity, the iNC may be physically co-located with the RANaaS entity.

iJOIN Local Gateway (iLGW): Logical entity implementing a subset of the functions of a P-GW. It is logically connected with a (v)eNB, but can be physically located anywhere in the RAN.

iJOIN Transport Node (iTN): Physical entity located between iSC and RANaaS, or between RAN and core network. Each iTN is essentially a transport node operating at a different protocol stack layer depending on the particular functional split, and a set of iTNs is forming a backhaul network whose forwarding plane can be configured by an iNC

4 iJOIN System Concept

4.1 Key Enablers

Future mobile networks will have to provide an exceptionally greater traffic volume with diverse data rates from machine-to-machine (low data rates) to 3D applications (high data rates) [27]. The four main drivers of this development are an increasing number of user terminals, more diverse mobile services and terminals, more data-rich internet content, and more powerful devices. Beside the pure quantitative requirements to future mobile networks, also more qualitative requirements are applied. Future mobile networks are expected to be flexible, programmable, and virtualised. Flexibility allows for integration of heterogeneous technologies and services. Programmability allows for simpler network evolution and integration of novel technologies. Virtualisation allows for multi-tenancy improving the utilisation of network resources.

In order to achieve these quantitative and qualitative goals, iJOIN considers the two key enabling technologies *small-cell base-stations* and *centralised processing*. The use of very dense, low-power, small-cell networks is one of the most promising options to increase the system throughput. They exploit two fundamental effects. Firstly, the distance between the radio access point and users is reduced and the data rate increases super-linearly by the inverse of the distance. Secondly, the spectrum is used more efficiently because each radio access point uses the same frequency resources. In the best case, this offers a linear increase of system capacity in the number of deployed access points. Small cells complement existing macro-cellular deployments which are still required to provide coverage for fast-moving users and in areas with low user-density. By contrast, small-cells allow for providing data traffic where the actual demand arises.

While this may improve the data rates, it complicates network planning and deployment, backhaul provisioning, and network maintenance. As networks become denser, inter-cell interference increases and interference scenarios become more complex due to multi-tier interference. Furthermore, the higher the deployment density is, the higher is the chance that a certain radio access point will carry no traffic or only a low traffic-load due to spatial and temporal traffic fluctuations. Currently, 15-20% of all sites carry about 50% of the total traffic [28]. This implies that a considerable number of sites consume energy and computational resources, even though they carry no traffic or only a negligible amount of traffic.

Centralised processing permits the implementation of efficient interference avoidance and cancelation algorithms across multiple cells. It provides the means to selectively turn Radio Access Points (RAPs) on and off in order to load-balance traffic in scenarios with high traffic fluctuations. Centralised-RAN (C-RAN) is one possible way to efficiently centralise computational resources, to balance throughput fluctuations, and to implement inter-cell coordination. Furthermore, C-RAN is an important architecture to enable fast service creation, to support reprogrammable and flexible radio interfaces, to virtualize radio access networks, and to allow for flexible and elastic usage of resources [29][30].

In C-RAN, multiple sites are connected to a central data centre where all baseband processing is performed. Transmitted and received radio signals are exchanged over fibre transmission lines (called front-haul) between Remote Radio Heads (RRHs) and the data centre. At present, only fibre-links or mmWave-links [31] are capable of supporting these data rates. This constitutes the main drawback of C-RAN, which is the need for a high-capacity and low-latency front-haul links. Due to the use of either optical fibre or mmWave-links, C-RAN deployments are less flexible and very expensive if the necessary infrastructure does not already exist. Hence, there is a trade-off between centralised processing requiring high-capacity front-haul links, and de-centralised processing which can be supported by traditional backhaul.

4.2 Key Concepts

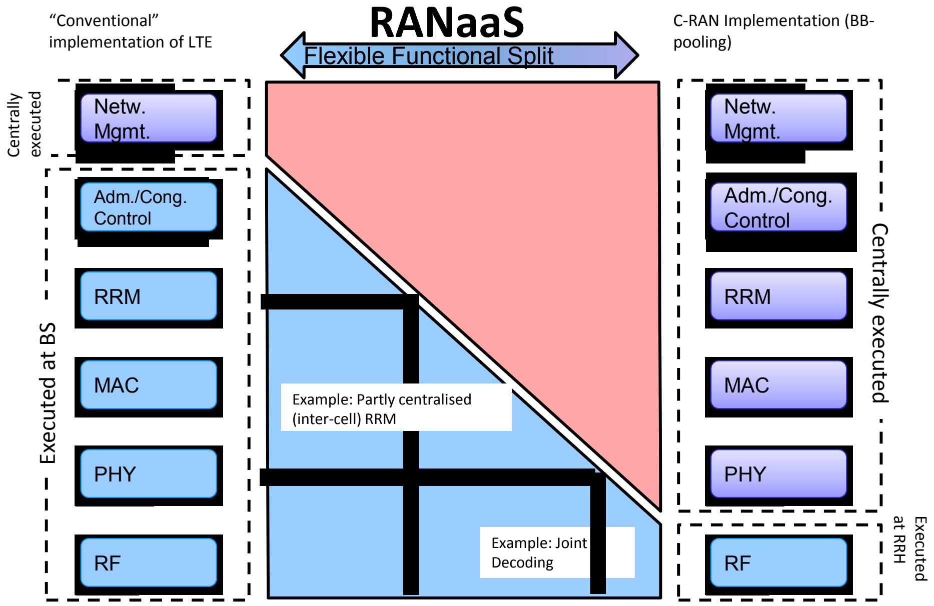

Figure 4-1: iJOIN key concept RAN as a Service

iJOIN introduces two novel key concepts which allow for applying centralised processing to small-cell networks. The first key concept is RAN-as-a-Service (RANaaS), in which radio access network (RAN) functionality is *flexibly centralised* through an open IT platform based on a cloud infrastructure. The basic concept is illustrated in Figure 4-1. The left side of Figure 4-1 exemplifies a traditional LTE implementation where all functionality up to Admission/Congestion Control is locally implemented at the Base Station (BS). The right side illustrates the C-RAN approach where only the radio front-end (RF) is locally implemented and all other functionality is centralised, including digital baseband processing. By contrast, RANaaS does not fully centralise all functionality, but rather flexibly centralises part of the RAN functionality and offers this as a service. iJOIN introduces different preferential functional splits, and it analyses and quantifies their benefits and requirements.

RANaaS is an application of the XaaS-paradigm that says that any kind of service may be centralised by a cloud-platform. Services are provided on demand. Resources are scalable, can be better controlled and optimised, and may be pooled independently of the location and transparently to the user. In iJOIN, an implementation on virtualised cloud-infrastructure has been investigated in order to determine the feasibility of the RANaaS concept. iJOIN further quantified the throughput performance gains obtained by centralised processing and computational diversity gains in a RANaaS deployment.

All three network domains (RAN, backhaul network, and data centre) impose different requirements and ~~provide~~ have different characteristics. Hence, in order to select and operate a functional split option, it is necessary to appropriately control and monitor all three network domains. This is illustrated in Figure 4-2 which shows the three network domains applied to the iJOIN architecture (which is detailed in Section 5). Using the RANaaS concept the functionality of a BS is split between the physical iJOIN small cell (iSC) and the upper part of the virtual eNB (veNB) which is executed on the RANaaS platform. The centralisation of functionality is mainly determined by the backhaul capacity and end-to-end latency between iSC and RANaaS instance. Hence, the system must keep track of the backhaul characteristics and maintain minimum requirements in order to centralise RAN functionality.

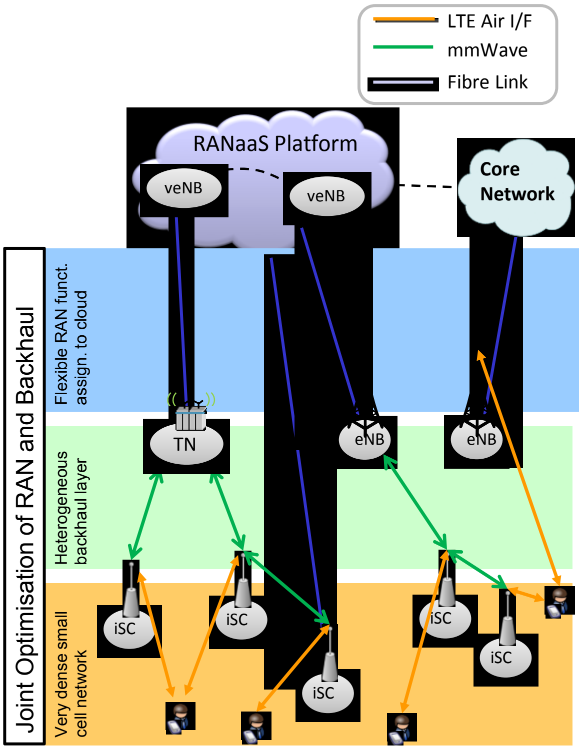

Figure 4-2: iJOIN key concepts

Consequently, iJOIN's second key concept is the joint design and operation of RAN, backhaul network, and RANaaS platform by means of Software-Defined-Networking (SDN). iJOIN developed novel operation and management algorithms to optimally operate a very dense mobile network, to flexibly split functionality between iSC and RANaaS instance, and to jointly operate the radio access and backhaul network layer. The main objective is to optimise the RAN system throughput and provide services instantly and efficiently in cost, energy, complexity, and latency wherever and whenever the demand arises. Both key concepts are explained in further detail in Section 6 and Section 7.

4.3 iJOIN System Design Criteria

In this section, we take a top-down approach to the iJOIN architecture and system design. We start from the main objectives of the project and identify the inherent trade-offs between the different objectives. Building on our trade-off analysis and several considerations on the objectives, we devise some system design criteria that will guide the design of the different candidate technologies of the iJOIN architecture and ensure the consistency of the individual components to achieve the overall system goals.

iJOIN project objectives

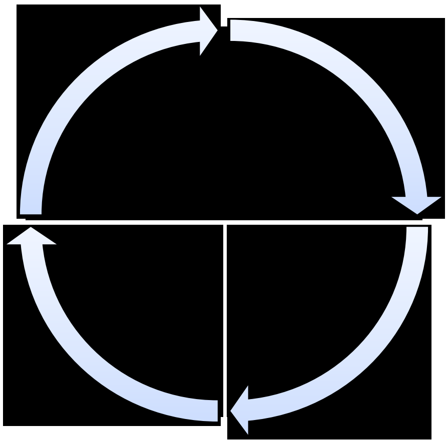

Figure 4-3: iJOIN's four key objectives

As illustrated in Figure 4-3, iJOIN targets four objectives of particular interest:

- ***Area throughput***: iJOIN seeks to increase system area throughput through an improved system efficiency, i.e., it targets to increase the system throughput per covered area within the same spectrum band.
- ***Energy efficiency***: Another key objective of iJOIN is to improve the network-level energy-efficiency, measured as an Energy Consumption Index (ECI) (in Joule/bit), i.e., improve the relative energy-consumption per bit of the network. Indeed, even though the *overall* energy-consumption may increase due to the densification of the network, the goal is to avoid an increase of energy consumption with the same slope as the system-throughput increases.
- ***Cost efficiency***: iJOIN also aims at reducing the cost of deploying and operating the mobile network, such that the cost-per-bit is reduced. This applies to capital expenditures (CAPEX) as well as operational expenditures (OPEX).
- ***Utilisation efficiency***: iJOIN intends to increase the utilisation efficiency in relevant scenarios by concentrating traffic where it is actually needed instead of ubiquitously and continuously providing peak throughput. This involves not only utilisation of the radio access resources, but also of the backhaul resources as well as the computational resources.

In principle, each candidate technology introduced in iJOIN project aims at optimising one particular objective of the above list. However, in addition to the target objective, each candidate technology may also impact other objectives. Therefore, co-deploying multiple candidate technologies could potentially result in a set of contrary effects on multiple objectives unless a careful architecture design is conducted to avoid this. The scope of this section is to evaluate these effects and ensure that different candidate technologies are deployed to jointly achieve the overall goals of the project. To achieve this, we next present an analysis that aims at jointly optimising the metrics of interest in the project taking into account the trade-offs between the different objectives as well as the interactions between the different candidate technologies.

Trade-off analysis and considerations

Before providing the overall system guidelines, we first analyse the inherent trade-offs involved in the various metrics of interest. The analysis here is at a high level, without considering the specific candidate technologies investigated in the iJOIN project and their effects onto each one of the metrics – the reader is referred to the end of this section for a candidate technology specific analysis.

The following bullets give an overview of the existing trade-offs for each of the four metrics listed before:

- There is a trade-off involved between area throughput and energy efficiency. For instance, the more nodes we switch off in the network, the better the energy efficiency, but this could come at the cost of reducing the maximum offered area throughput. This trade-off, also traditionally known in literature [42] is illustrated in the ideal case by Figure 4-4:, which shows the region of feasible operating points in terms of energy efficiency and area throughput performance for each of these points. In order to optimise performance, we want to be in the edge of that region. However, while being at the edge, we have to trade off throughput with energy: as it can be seen in the figure, the better the maximum offered area-throughput, the lower the energy efficiency, and vice versa.

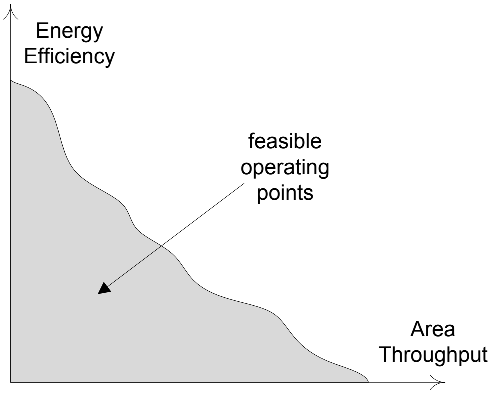

Figure 4-4: Trade-off of energy-efficiency and area-throughput

- The utilisation metric is considered as a side-effect. While this is an important metric as it reflects the efficiency of the operating point of the system, it is not an objective in itself for the design. For instance, when considering the utilisation in the RAN, the actual metric that we want to optimise is the area throughput, while when considering the efficiency of computational resources the actual metric to optimize is cost.
- In the general case in which the capacity of the air interface of a small cell is not equal to the corresponding backhaul capacity, the utilisation level at the air interface and at the backhaul differ, and therefore there is a trade-off between the utilisation of these two resource types. On the other hand, for a given traffic demand the utilisation of computational resources is orthogonal to the other operational parameters, i.e. air interface and backhaul, and it can be optimised independently.
- Cost efficiency in terms of capital expenditures (CAPEX) is performed at a planning stage, and hence does not have any effect on the other objectives which are addressed at a different level (namely at an operational stage). It is worthwhile noting that there is some relationship between this parameter and the utilisation of computational resources. Indeed, the more efficient the architectural design for centralised computation, the lower the CAPEX and at the same time the higher the utilisation of computational resources. However, utilisation here is rather a way of evaluating the efficiency of the system design, than a trade-off or guideline that has an impact on the system design itself.
- As far as cost efficiency in terms of operational expenditures (OPEX) is concerned, our assumption is that OPEX is mainly dominated by energy consumption. Indeed, while there are other OPEX components that are also very significant, such as rental costs and maintenance, these are rather unrelated from the rest of the operational metrics and therefore can be optimised independently. As a result of this assumption, by finding a good trade-off between area throughput and energy efficiency, we are also optimising OPEX.

System design guidelines

The above considerations show that some of the objectives are orthogonal and therefore can be independently optimised. For other objectives there exists some trade-off and therefore we need to choose how to combine these objectives and which one to prioritise. In the following, we provide several system guidelines that illustrate the choice that has been taken by the project in each of these cases. These guidelines are used as the basis for the design of the various candidate technologies within the project architecture.

- *System design guideline #1:* ***Area Throughput***. In the iJOIN architecture, area throughput is considered the performance objective of highest priority as it is the one that reflects user satisfaction. In order to make sure that this objective is met, the iJOIN architecture defines a certain throughput demand, which is considered to ensure user satisfaction, and guarantees that this throughput demand is met by the architecture.
- *System design guideline #2:* ***Energy efficiency***. Next to area throughput, the second highest priority objective in iJOIN is energy efficiency as this is one of the main drivers for OPEX from the architectural design standpoint. In order to optimise this metric, the choice made by the iJOIN architecture is to operate at the point that minimises energy consumption while guaranteeing the desired throughput demand. That is, from all the possible operating points that satisfy the required throughput demand, we choose the one with the smallest energy consumption. This choice is illustrated in Figure 4-5:.
- *System design guideline #3:* ***Utilisation efficiency***. Utilisation is of lower priority than area throughput and energy efficiency as it does not directly contribute to increasing user satisfaction (area throughput) at the lowest cost for the operator (energy efficiency). However, it is in general desirable to operate at a point that utilisation across small cells is as evenly distributed as possible, as operating close to the maximum capacity risks harming area throughput in case of unpredicted deviations from the operating point. Based on this, in iJOIN we chose, from all possible operating points that guarantee throughput demands and minimise energy consumption, the one that presents the best balance in terms of utilisation. As for air interface vs. backhaul resources, we prioritise balancing air interface resources, as these are typically less predictable.

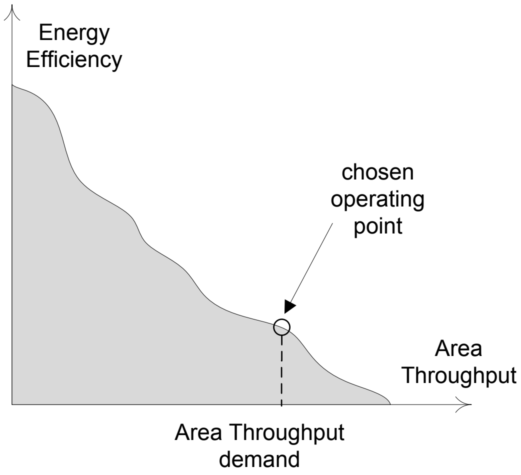

Figure 4-5: Choice of operating point

- *System design guideline #4:* ***CAPEX***. Cost efficiency in terms of CAPEX is addressed by optimising the existing trade-off between centralised computation (which reduces the capital expenses in computational resources) and backhaul capacity (which increases capital expenses in backhaul deployment). This guideline is orthogonal to the previous ones as it relates to the planning phase and therefore is independent from the architectural operation.

Impact of candidate technologies

In the following, we analyse the impact that each of the candidate technologies has on each of the objectives defined above. Indeed, while some of the project's candidate technologies only impact one of the objectives and do not have to consider the various trade-offs involved, others do have side effects that need to be considered. This is illustrated in the following tables, which show the main objectives for each of the candidate technologies as well as the other objectives which are potentially affected. A full description of these candidate technologies is given in deliverables D2.2 [4], D3.2 [10], and D4.2 [13].

Table 4-1: Candidate Technologies in WP2 and their mapping to iJOIN objectives

CT	Area Throughput	Energy Efficiency	Utilisation Efficiency	Cost Efficiency
2.1 - In Network Processing	++	+	+	0
2.2 - Multi-Point Turbo Detection	++	0	+	0
2.3 - Joint Network-Channel Coding	++	0	+	0
2.4 - DL CoMP with backhaul constraints	++	+	0	0
2.5 - Inter-Cell Interference Coordination	++	+	0	0
2.6 - Data Compression over RoF	+	0	+	+
2.7 - Millimetre Wave Backhauling	++	+	0	+

Table 4-1 shows the candidate technologies (CTs) that are investigated in WP2 and their impact on the four objectives. Apparently, all CTs except for CT 2.6 address mainly area throughput, i.e. under a given set of enabled iSCs, these CTs maximize the achievable throughput. Therefore, these CTs extend the feasible operating region along the area throughput axis. However, this increased area throughput may be at the cost of additional energy consumption. In uplink, CT 2.1 introduces in-network processing, CT 2.2 introduces multi-point turbo-detection, and CT 2.3 introduces joint network-channel coding. The former two CTs introduce additional backhaul traffic. Similarly in downlink, CT 2.4 and 2.5 investigate cooperative multi-cell processing which optimise the degree of cooperation. However, there is also a requirement for additional backhaul traffic. Finally, CT 2.7 optimises achievable data rates in RAN and BH jointly and therefore does not require additional backhaul traffic but only additional processing at the RANaaS instance. The additional backhaul traffic implied by CT 2.1-2.5 may cause an increased energy consumption in the backhaul. However, the main source of energy consumption for backhaul equipment is fixed energy which does not scale with the actual backhaul traffic [19]. Hence, we expect that the increase of energy consumption to apply cooperative technologies will be rather marginal compared to non-cooperative technologies. Hence, during the system design step where a data rate constraint is decisive, an operating point with lower energy consumption (higher energy efficiency) may be chosen. The same applies to utilisation efficiency.

The only exception to the above description is CT 2.6 which reduces the required data rate on the fronthaul link and therefore represents rather an enabling technology for RANaaS and contributes to the CAPEX of the system because fronthaul links at lower capacity can be chosen.

Table 4-2: Candidate Technologies in WP3 and their mapping to iJOIN objectives

CT	Area Throughput	Energy Efficiency	Utilisation Efficiency	Cost Efficiency
3.1 - Backhaul Link Scheduling	++	0	+	0
3.2 - Cell Selection	++	0	+	0
3.3 - Energy-Efficient RRM at Access and Backhaul	0	++	0	+
3.4 - Semi-Deterministic Scheduling	++	0	0	+
3.5 - Inter-Cell Interference Coordination in RANaaS	++	0	+	0
3.6 Utilisation efficiency, joint RAN/Cloud scheduling	+	0	++	+
3.7 - RRM for Scalable Multi-Point Turbo Detection	++	0	+	0
3.8 - RRM for In-Network Processing	++	0	+	0
3.9 - Multi-level scheduling	++	0	+	0

Table 4-2 shows the candidate technologies that are investigated in WP3 and their impact on the four objectives. Again, we can expect side-effects on other objectives. CT 3.1 optimises backhaul link scheduling to increase throughput but thereby also improves the utilisation of available backhaul resources. Similarly, CT 3.2 optimises cell selection based on joint information from radio access and backhaul network. CT 3.4 maximises the area throughput through backhaul-latency aware link adaptation and scheduling. CT 3.5 introduces a novel inter-cell interference coordination scheme which improves the area throughput and as a side-effect the RAN utilisation. CT 3.6 improves the utilisation of computational resources in the RANaaS, avoiding computational outage and thereby, as a main side effect, improving area throughput for a given computational resource constraint. CT 3.7 and CT 3.8 are the matching MAC protocols for CT 2.2 and CT 2.1, respectively. Finally, CT 3.9 introduces a multi-level inter-cell interference coordination which mainly improves the area throughput and, as a side-effect, utilisation of available radio resources. Similarly to the candidate technologies of WP2, we observe that all CTs (except for CT 3.3) improve the area throughput for a given set of available backhaul and radio access resources. As the area throughput improves, also the utilisation of the resources is improved.

The only exception is CT 3.3 which improves the energy-efficiency by optimising the activity pattern of iSCs on a very short timescale under given quality of service side-constraints, i.e. delay and throughput. Hence, we define at first the area throughput constraint and based on this constraint, CT 3.3 would optimise the energy efficiency metric.

Table 4-3: Candidate Technologies in WP4 and their mapping to iJOIN objectives

CT	Energy Efficiency	Area Throughput	Utilisation Efficiency	Cost Efficiency
4.1 - Distributed IP Anchoring and Mobility Management	0	0	++	0
4.2 - Network Wide Energy Optimisation	++	0	0	0
4.3 - Joint Path Management and Topology Control	0	0	++	++
4.4 - Routing and Congestion Control Mechanisms	0	0	++	+
4.5 - Network Wide Scheduling and Load Balancing	0	0	++	0

Table 4-3 shows the candidate technologies that are investigated in WP4 and their impact on the four objectives. In contrast to WP2 and WP3 CTs, WP4 does not directly address area throughput, but mainly utilisation, cost and energy efficiency. CT4.1 improves utilisation efficiency by properly choosing where to anchor user traffic on a per-flow basis, taking into account the network status. CT4.2 optimises the energy efficiency by jointly considering the radio access and the backhaul network while satisfying the user demands. Besides, this optimisation is performed on a larger timescale than CT3.3. CT4.4 and CT4.5 optimise mainly utilisation efficiency by performing congestion control and load balancing on the backhaul. Finally, CT4.3 improves utilisation efficiency by optimising the number and placement of the RANaaS entities. By doing so, cost-efficiency is also improved as the number of RANaaS entities is minimum to meet a given traffic requirements.

Candidate technologies' design

The guidelines defined earlier in this section, along with the impact and trade-off analyses conducted for the candidate technologies, provide a high level perspective of the objectives of the iJOIN architecture, and how the different objectives are traded off in the architecture design. We have also analysed the impact that each candidate technology has on the objectives.

The remaining challenge is to map the system-level design guidelines discussed above onto the specific project's candidate technologies. In particular, we need to translate the general guidelines to candidate technology specific guidelines that can be used in the design of each of the candidate technologies comprised in the iJOIN architecture. Such guidelines are derived from the general ones, this ensures that the design of each candidate technology fits within the overall framework and contributes to the design of a harmonized and well integrated system.

Figure 4-6 shows the candidate technology's specific guidelines obtained by performing the above mapping. As it can be seen from the table, many of the candidate technologies solely focus on optimizing throughput performance, either in the access (CTs 2.1-2.5, 3.1, 3.4 - 3.9) or the backhaul (CTs 2.6 - 2.7), and hence we do not need to consider any trade-off in their design. Two of the candidate technologies (CT 3.3 and 4.2) focus on energy efficiency, and need to ensure that Area Throughput demands are satisfied when minimizing energy consumption. Finally, the remaining candidate technologies focus on optimizing utilization, either in the backhaul (CTs 4.1 - 4.2, 4.4 - 4.5) or in the access (CT 3.2). These candidate technologies take as constraints area throughput and energy efficiency, and optimize utilisation within these constraints.

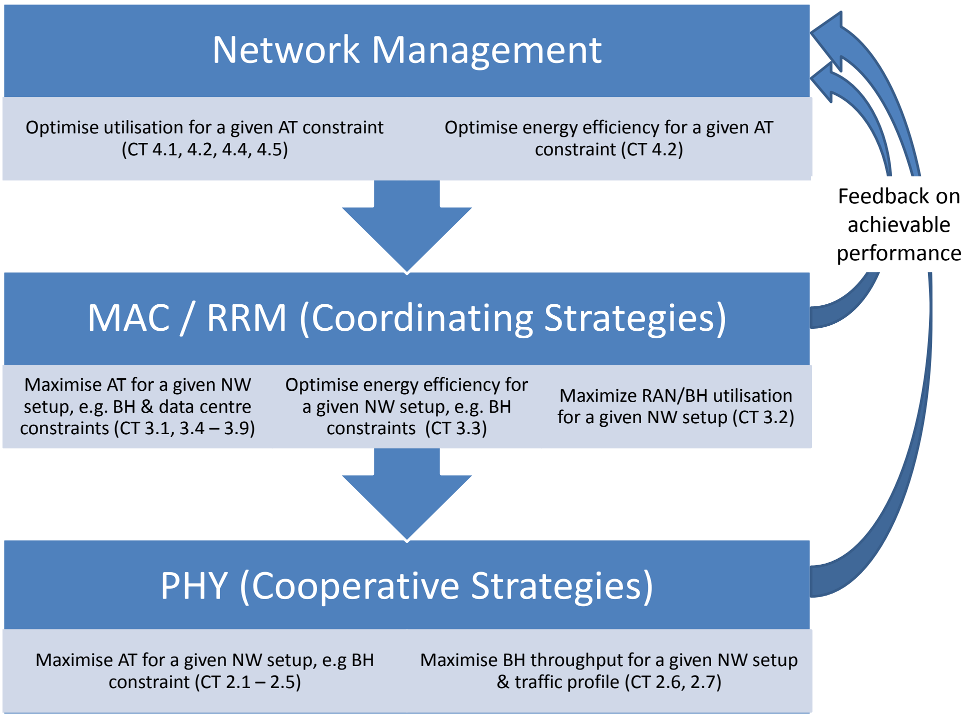

Figure 4-6: CT-specific design guidelines resulting from project design guidelines

5 Final iJOIN Architecture Definition

This chapter first presents an overview of the baseline architecture considered as reference in the iJOIN project. Then, it discusses the final iJOIN logical architecture by defining all entities relevant for the project. In addition, we describe the cloud-architecture and its relation to ETSI NFV (Network Function Virtualization) [26]. After that, we provide the final iJOIN functional architecture which lists all iJOIN modules and interfaces involved, both from a project-wide viewpoint as well as from individual work package viewpoints. Finally, the iJOIN Physical Architecture is presented with a short description of all four common scenarios.

5.1 Baseline Architecture

The iJOIN project provided an evolutionary path of the current 3GPP architecture for LTE/LTE-A which will support the innovations developed within this collaborative project. For that purpose, a baseline architecture has been chosen in order to clearly identify the novel entities introduced by the project and also as reference for evaluations performed by iJOIN.

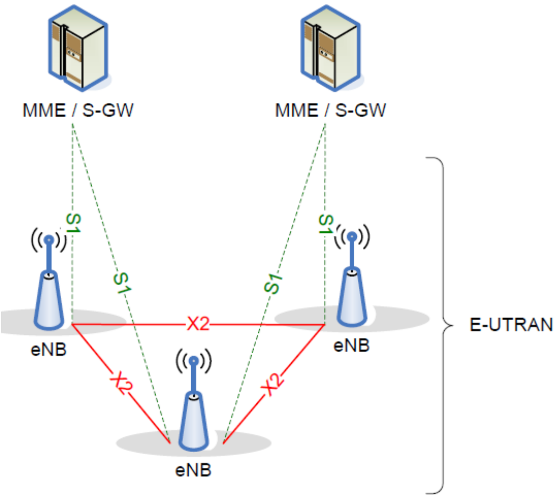

Figure 5-1: LTE Rel.10 Logical Architecture

In particular, LTE/LTE-A Release 10 [6] has been chosen as baseline architecture. Figure 5-1 shows the reference logical architecture of 3GPP Evolved Universal Terrestrial Radio Access Network (E-UTRAN), where evolved NodeBs (eNBs) are connected with each other through X2 interfaces and to the Evolved Packet Core (EPC) elements through S1 interfaces.

The mobility management entity (MME) and serving gateway (S-GW) are the connecting points of the E-UTRAN to the EPC; a detailed description of both elements and S1 and X2 interfaces can be found in D5.1 [1]. It is important to clarify that the innovations introduced by iJOIN maintain these legacy interfaces and elements. iJOIN introduces additional logical entities and interfaces that permit to increase network performance, give more flexibility and enable an efficient deployment of evolved mobile networks.

5.2 iJOIN Logical Architecture

In the past, the trend was to push the computation burden toward the last miles in order to reduce the round trip time and improve the reactivity of the system, e.g., ARQ vs HARQ. With dense small cell deployments being a promising solution to answer the growing need of capacity, (partial) centralisation is required to deal with complex interference situations. To cope with upcoming dense deployments of LTE-based small cells, iJOIN proposes the evolutionary architecture given in Figure 5-2 which is a refined version of the proposal introduced in D5.1 [1]. Most notably, several changes have been applied in the way the virtual base station concept is implemented within the cloud architecture.

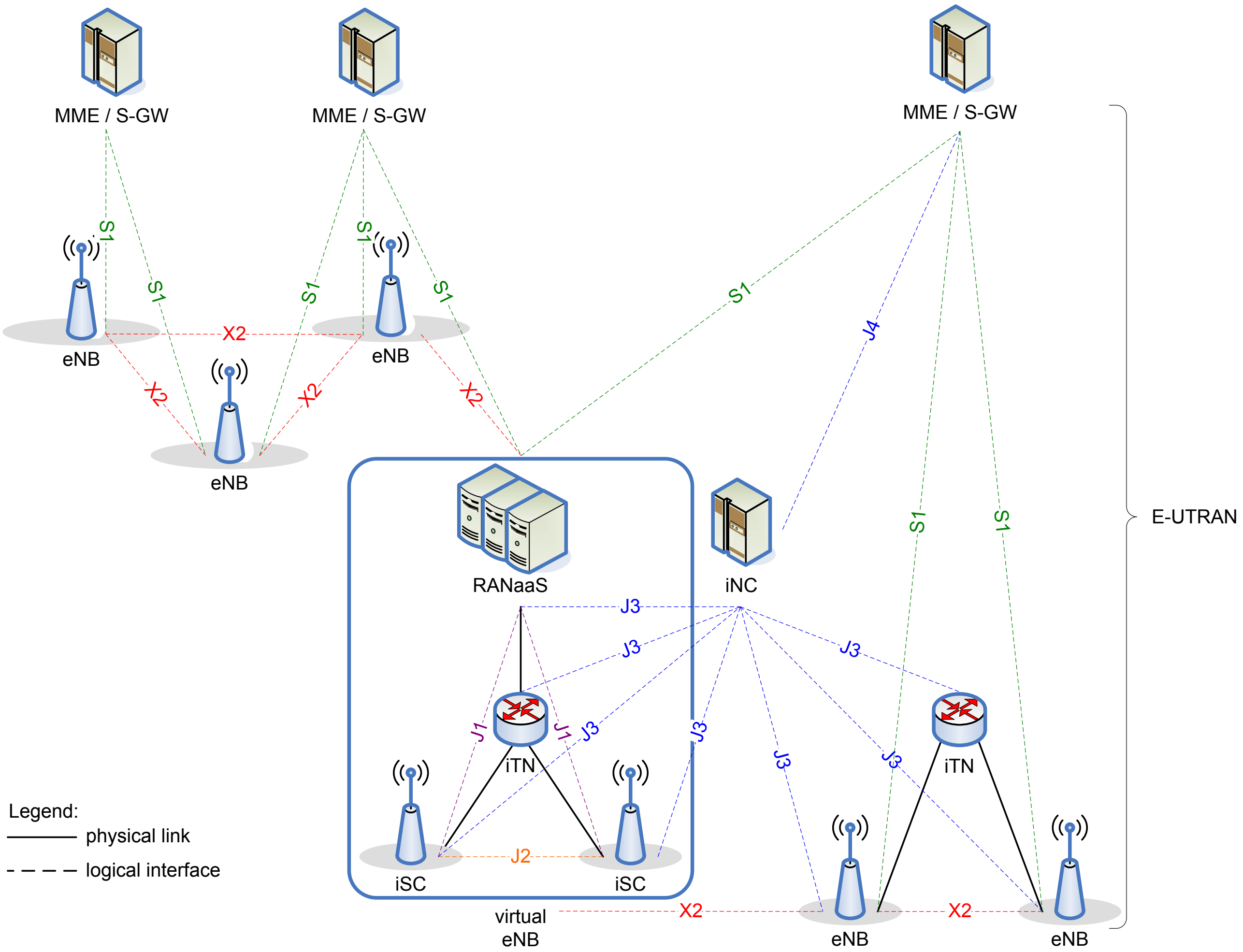

Figure 5-2: iJOIN Logical Architecture

Based on the architecture shown in Figure 5-2, the RANaaS and iSCs entities appear as standard-conform eNBs to the existing network. Therefore, such virtual eNB (veNB) entity can be seamlessly integrated in existing architectures. The core network does not need to know that the RAN functionalities are effectively split between iSCs and RANaaS. It only needs to know the routes for user and control planes, which by default will be terminated at the RANaaS entity.

One veNB comprises the RANaaS instance running on a cloud platform (veNB upper domain) and one or several iSCs (veNB lower domain). Within one veNB, the iSCs and the RANaaS instance are connected through the J1 interface, while the iSCs can exchange information directly with each other using the J2 interface. Comparable to a legacy eNB, one veNB can setup one X2 connection with other (v)eNBs supporting the exchange of standardised 3GPP signals.

In order to optimise jointly the RAN and the backhaul, an iJOIN Network Controller (iNC) entity is introduced. It configures the routing among the backhaul Transport Nodes (iTNs) based on configurable constraints, e.g., RAN/backhaul load, user density, or mobility pattern. This SDN-based controller solution can be applied to the proposed RANaaS/iSCs setup as well as to legacy LTE deployments. The iNC relies on the J3 interface connecting each of the involved entities as depicted in Figure 5-2. In addition, the iNC is also able to dialog with the core network through the J4 interface for the purpose of routing, anchoring, and mobility.

In the following subsections each logical entity will be more precisely described.

5.2.1 RANaaS Cloud Architecture

The RANaaS component is designed and implemented by leveraging a “standard” cloud computing technology baseline. The RANaaS implements an IaaS system, deployed and executed over general purpose, industry standard computational nodes (servers) with special hardware extensions where necessary. The flexible

RAN functional split allows shifting some RAN functions to the centralised RANaaS entity without imposing mandatory constraints on the cloud computing infrastructure or its software technology.

The RANaaS cloud architecture is shown in Figure 5-3. The entity denoted by RANaaS corresponds to a *RANaaS instance*, which is an actual implementation of a *RANaaS platform*. It involves the combination of a physical computational storage network infrastructure with a set of software modules providing the cloud management functions.

The *RANaaS platform* denotes the technology baseline implementing one or more specific RANaaS instances. For example, in the case of cloud IaaS based RANaaS, the RANaaS platform encompasses the whole physical infrastructure resources, e.g., servers, storage, network interconnecting appliances and cloud management software (e.g., OpenStack/KVM) to create, run, manage and decommission virtual machines and hypervisors.

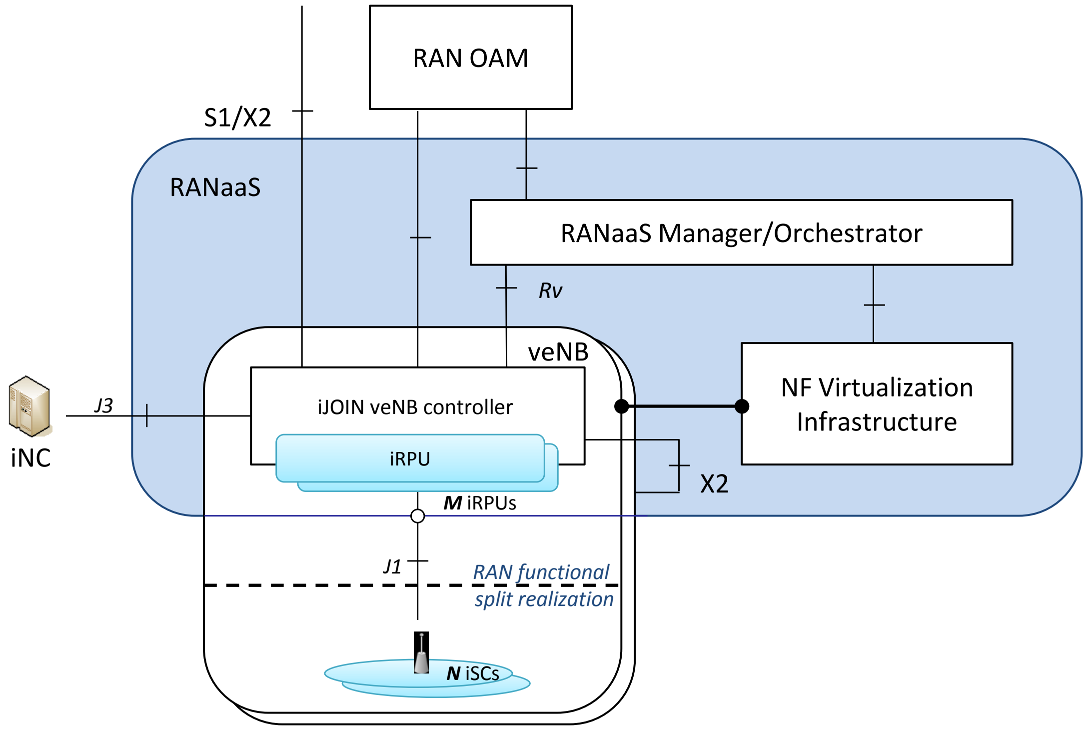

Figure 5-3: RANaaS Cloud Architecture

Components within the RANaaS platform comprise functions or functional blocks related to the RAN operation, i.e., the veNB, the iJOIN veNB Controller (iveC), iJOIN virtual RAN Processing Units (iRPU), and management and orchestration functions including the RANaaS manager and the Network Function Virtualization Infrastructure (NFVI). Note that the latter could be an instantiation of the ETSI NFV framework, such that the RANaaS manager uses (or integrates into) NFV management and orchestration and NFVI reference points [24]. Network functions such as the veNB controller can be therefore seen as Virtual Network Functions (VNF) according to this framework. For further details see Section 5.2.1.5.

Table 5-1: Function placement in RANaaS platform

	iveC	RANaaS manager	iRPUs
veNB lifecycle management		X	
veNB performance monitoring		X	
Triggering of veNB scaling		X	
veNB bootstrapping		X	
J1 interface setup	X		
S1/X2 interface termination	X		
GTP tunnel management	X		
Intra-veNB RRM and RRC	X		
RAN functional split management	X		
veNB fault monitoring	X		
LTE upper protocol stack according to functional split			X

In the following, RANaaS network functions are defined in more detail. Table 5-1 provides an overview of the functional assignments between the different RANaaS platform components.

5.2.1.1 Virtual eNodeB (veNB)

The RANaaS architecture requires that the functionality of a "classic" eNB can be decomposed into re-assignable RAN functions, where each function can be assigned either to the RANaaS instance or the iSC. There are four main characteristics of a veNB:

Functional decomposition: the functionality executed by a veNB can be decomposed into modules. Each module implements a functionality of the RAN protocol stack such as PHY layer, MAC layer, or RRC layer procedures.

Functional split: each of the modules may be located either locally at the iSCs or centrally at the RANaaS instance. Furthermore, only a subset of all modules may be executed at the iSCs while another subset is implemented at the RANaaS instance. The interaction between modules, the placement of modules, and the interaction of modules with other logical entities is controlled by the veNB Controller.

Functional interworking (Control Plane): interfaces that are used for the coordination and control of functional split in order to transparently configure which modules are executed at the iSC and which modules are executed in the RANaaS instance.

Logical unity: on a black-box view, e.g. from the core network perspective, the veNB appears as a 3GPP LTE eNB in order to ensure compatibility with 3GPP specifications.

A veNB comprises one iJOIN virtual eNB Controller, a set of iSCs, and a number iRPUs, which are executing functions of the upper layer of the LTE protocol stack according to the configured functional split. The interworking of iRPUs and iSCs therefore constitutes the realisation of a specific functional split. The split is, in our assumption, specific to a virtual eNodeB to ensure that within the logical network entity, the functional components can be managed and addressed coherently. This is in line with the assumption that a virtual eNodeB can be mapped to a geographic area with a specific backhaul deployment such that the need for different functional realisations within a veNB is unlikely.

5.2.1.2 iJOIN Virtual eNB Controller (iveC)

The iveC is the interface termination entity of the veNB towards the RANaaS manager and the EPC network. It is responsible for various control and management tasks but also for implementing a compatible interface to existing networks. In order to benefit from centralised RAN functions, inter-iSC cooperation is necessary. This is implemented by the iveC across the iRPUs under its control, e.g. in terms of centralised radio resource allocation or connection control. The exact scheme depends on the implemented RAN functional split; for more details see Section 5.2.1.5.

The iveC is also in charge of controlling the RAN functional split. The veNB controller and the iNC determine the feasible functional splits based on the backhaul topology and access technologies, e.g., fibre, wireless point-to-point, or wireless point-to-multipoint [14], the network status, the RANaaS instance placement, and the user data rate demand. This is necessary because not every functional split is possible for a given network scenario as the backhaul network delay and bandwidth might not be sufficient to centralise a given function. Note that the functional split control should also take into account the capabilities of the iSCs, e.g., CPU processing power, availability and type of antennae, and the level of connectivity among them, i.e., the expected latency and bandwidth that can be obtained between neighbouring iSCs, and between the iSCs and the RANaaS instance.

The iveC also regularly monitors the state of the veNB, reacting upon events such as iSCs that become unreachable or present a faulty operation. Finally, the iveC is responsible for implementing the 3GPP eNB architecture as specified in [25]. This includes setup and termination of the eNB user and control plane interfaces (S1/X2), GTP tunnel management, and specific configuration.

5.2.1.3 RANaaS Manager

The RANaaS manager is responsible for performing management and orchestration tasks in support of veNB operation. For that purpose it maintains a management interface (denoted Rv in Figure 5-3) towards a set of veNBs or iveCs, respectively. The RANaaS manager performs lifecycle management, i.e. instantiation and decommissioning of veNBs, as well as performance monitoring and scaling operations. This needs to be a RANaaS-specific function since a veNB needs to fulfil carrier-grade operational requirements, equal to that of a dedicated eNB. The RANaaS manager interfaces towards the virtualised network infrastructure in order to request lifecycle and scaling operations.

Furthermore, the RANaaS Manager is responsible for the bootstrapping of veNBs. This encompasses several actions such as J1 interface establishment, information exchange with the RAN OAM, e.g. to request RAN configuration parameters, and setup of virtual network functions in the RANaaS cloud platform. More complex aspects could also be considered, for example in the case of sharing the infrastructure among multiple tenants. The bootstrapping should also take care of provisioning the iSCs with proper security credentials based on existing trust relationship, e.g., mutually trusted certificates. Note that an iSC lacking the required security credentials would not be validated and will be excluded from the veNB bootstrapping process.

5.2.1.4 iJOIN virtual RAN Processing Unit (iRPU)

The iRPU's main responsibility is the execution of the upper LTE RAN protocol stack according to the implemented functional split. This could be in its most extreme cases either the full RAN protocol stack (in case of a full centralisation), or only a proxy for Radio Resource Control (RRC) and Packet Data Convergence Protocol (PDCP) towards the iveC (in case of a full decentralization). In context of the RAN functional split options considered in iJOIN, we assume that at least PDCP and RRC are always executed in the RANaaS in order to obtain centralisation gains. It is assumed that the J1 interface is not directly terminated at an iRPU. Rather, a load balancing and routing function is in charge of distributing the data packets to their corresponding iRPUs or other network functions, e.g., the iveC for control plane messages.

5.2.1.5 Mapping to the ETSI NFV Framework

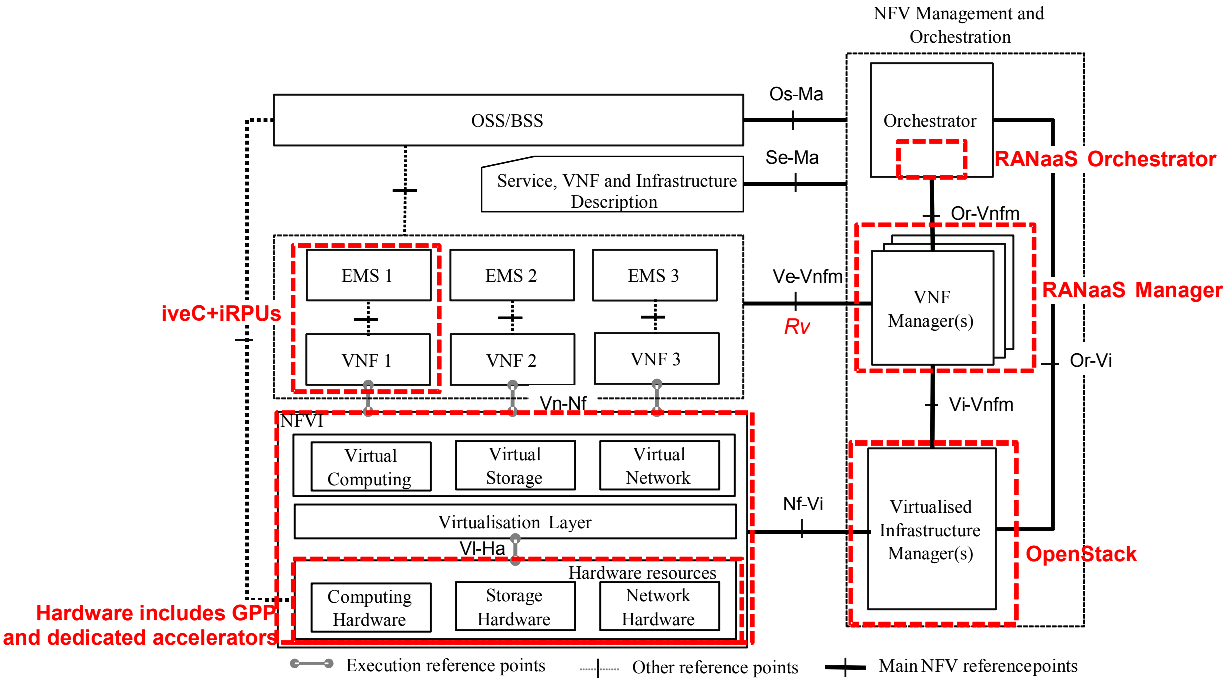

Figure 5-4: Mapping to ETSI NFV architecture

The ETSI ISG NFV is a working group which, since 2012, aims to evolve quasi-standard IT virtualisation technology to consolidate many network equipment types into industry standard high volume servers, switches, and storage. It enables implementing network functions in software that can run on a range of industry standard server hardware and can be moved to, or loaded in, various locations in the network as required, without the need to install new equipment. RAN virtualisation use cases are described, but not yet addressed in the current ETSI NFV specifications. The requirements and constraints in the RAN are still quite peculiar with respects to the case of intra- and inter-data centre network functions. To date, ETSI NFV is by far the most accepted NFV reference framework and architectural footprint. Hence, iJOIN analysed how the iJOIN architecture is aligned with the ETSI NFV model.

Figure 5-4 shows how the iJOIN RANaaS cloud architecture can be integrated into the ETSI NFV framework architecture as described in [24]. Five main functional blocks can be distinguished:

- The NFVI, which includes hardware resources, a virtualisation layer and virtual computing resources. The hardware resources can also be dedicated hardware accelerators which could be useful for computational intensive RAN processing tasks such as forward error correction.
- The VNF and Element Management (EM) entities, which are executed on the NVFI. The veNB executed in the RANaaS platform can be realised as a VNF (see below).
- The NFV management and orchestration (MANO) system, which comprises an orchestrator for system (potentially, network) wide orchestration of software and hardware resources, and a VNF manager, which is responsible for VNF life cycle management, including instantiation, querying, scaling, and termination of VNFs. In the RANaaS cloud architecture, the RANaaS manager can be mapped to the VNF manager. Certain aspects may also be realised as an orchestration module, depending on the overall software architecture of the deployed solution.
- The Virtual Infrastructure Manager (VIM), which is the component directly interacting with the NFVI to create virtual machines and allocate physical resources. It can be identified as the core IaaS platform, OpenStack in the iJOIN case.
- Furthermore, support functions such as OSS/BSS are necessary for operational management and monitoring of the system. Here, iJOIN foresees an interface with the RAN OAM system in order to allow for management decisions, e.g. in the RANaaS Manager, which takes RAN OAM parameters and information into account (not shown in the figure).

Figure 5-5 shows an example how a veNB can be realised as a VNF. A VNF in the ETSI NFV framework contains several VNF components (VNFC) [26], which are connected via logical interfaces and as such are not necessarily executed on the same virtual (and physical) hardware. The entities of the RANaaS cloud architecture can be mapped to VNFCs. It is assumed that per veNB, a single iveC is instantiated, which is in control of several iRPUs instances. A load balancing and gateway function terminates the J1 interface. This function, which could also be realised as a VNFC, is in charge of routing RAN user and control plane traffic to the corresponding iRPUs. Note that the exact composition of the VNF depends on the concrete functional split configuration. Several iRPUs may be connected in series, e.g. for different protocol layers. The scaling interfaces shown in the figure depend on the internal architecture of the VNF and on the chosen scaling method, which can be triggered by the VNF itself, e.g. by the iveC, or by the RANaaS manager.

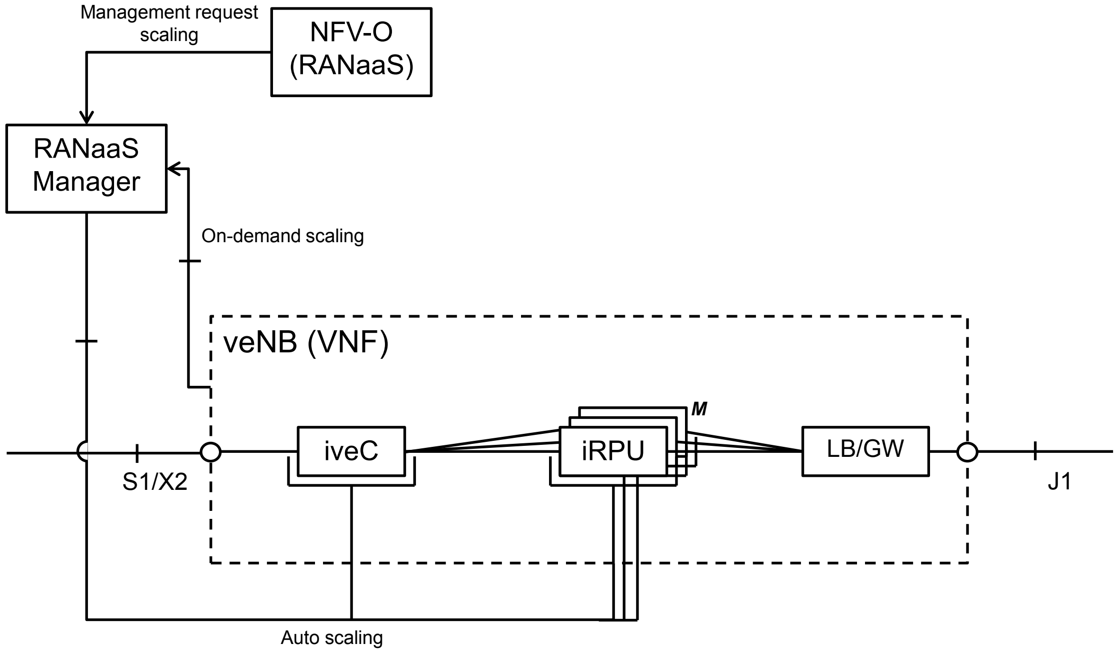

Figure 5-5: Realization of a veNB as VNF

5.2.1.6 Deployment and implementation aspects

RANaaS instances are physically hosted in a RANaaS PoP (point of presence), echoing a similar definition used in the NFV domain [24]. In the iJOIN context, a PoP is substantially equivalent to a data centre.

Several RANaaS instances can be configured and deployed on the same RANaaS platform and PoP. Every RANaaS instance must be correctly dimensioned for the parameters characterising the area it serves, i.e., areal extension, cell and user density, expected area throughput, and traffic profiles. Mechanisms need to be in place to guarantee a fully functional separation and isolation among the different RANaaS instances, protecting them by any possible interference from other co-located instances, or any other workload running in the same data centre. Such mechanisms are quite common in the data centre and cloud hosting domain, hence, no specific technology is needed to accommodate the RANaaS case.

A RANaaS instance is a cloud IaaS instance, physically deployed into dedicated nodes inside the PoP. Its core component is a standard IaaS management platform, e.g. OpenStack, on top of which iJOIN specific components are added. The NFVI can include particular hardware enhancements or capabilities to increase the virtualised function performance, which can be selectively used by the RANaaS for the virtual machines actually needing such performance surplus.

In the medium and long term, it is expected that the RAN virtualisation problem will tend to converge, and eventually to merge, with the more general NFV domain. iJOIN aims at providing an architectural footprint compliant with the most accepted NFV framework model, to guarantee that it can be evolved towards a full integration into NFV.

RANaaS will pose new challenges to data-centre architectures since they may not use existing data-centre platforms that were designed for internet services but rather dedicated platforms (which could still be called "commodity" due to the pervasiveness of mobile network technologies). For instance, the distribution and execution of processing jobs in data centres requires high-performance SDN architectures which route RAN data and address processing elements within data centres efficiently.

5.2.2 Virtual eNodeB Control Plane

The veNB control plane carries the signalling information that enables 3GPP LTE management functionalities such as security, connection control, mobility, interference coordination, and load balancing. Additionally, it enables the iveC to collect the information necessary for functional interworking. This information exchange is implemented through legacy S1 and X2 interfaces, which respectively connect a veNB to the MME and to neighbour veNBs. Moreover, the veNB communicates to the associated iSCs through the J1 interface.

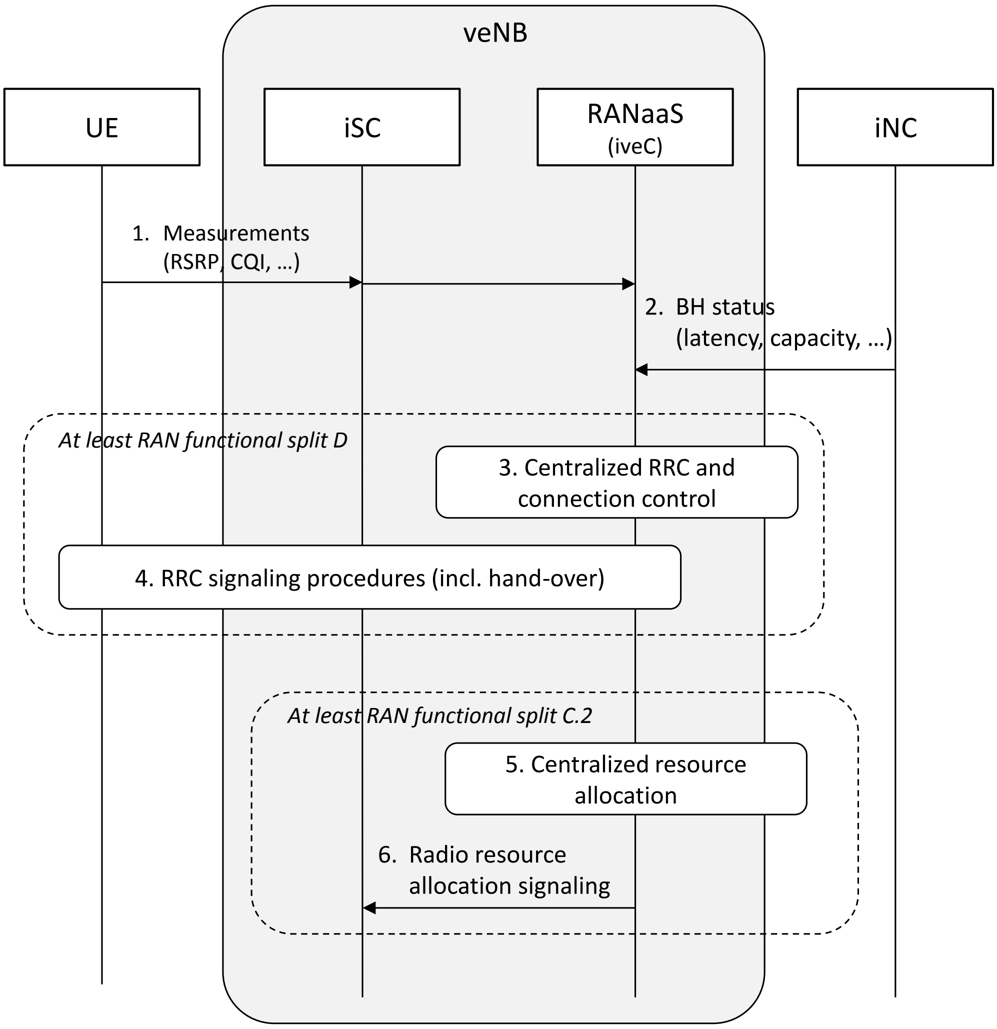

Figure 5-6: Message sequence chart for intra-veNB RRC and RRM

Figure 5-6 describes the message sequence chart supporting intra-veNB RRC and RRM. The selection of the best functional split is made at the iveC based on joint RAN and backhaul optimization. In particular, RAN measurements done at the UEs and the small cells are used at the veNB to capture the momentary status of the access network and system performance. On the other side, the transport network measurements realized at the iTNs are received at the iNC and then transferred to the veNB. By taking advantages of this feedback, the iveC can decided the functional split option supported by the transport network that results in the optimal improvement in terms of system performance (e.g., Area Throughput enhancement by reducing the inter-cell interference). Then, the iNC is informed about the decided functional split by the veNB and it adapts the transport network configuration such that the latency and bandwidth requirements are satisfied. Finally, also the access network is reconfigured and the selected functional split implemented.

5.3 Transport Network

The iJOIN transport network is realised by following a logically centralised approach based on SDN. This enables a very flexible backhaul network by optimizing its operation for the requirements posed by the different functional splits. The transport network architecture defined in iJOIN is shown in Figure 5-7 [14]. This architecture is made of three types of nodes:

- the iJOIN Network Controller (iNC),
- the iJOIN Transport Node (iTN), and,
- the iJOIN Local Gateway (iLGW).

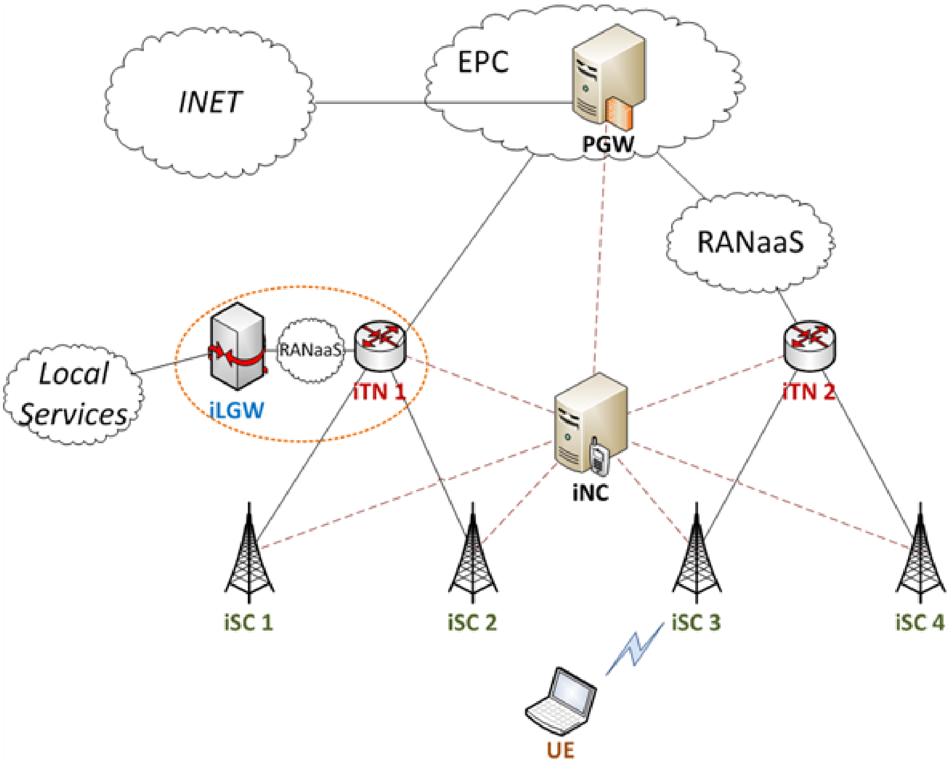

Figure 5-7: iJOIN Transport Network Architecture

5.3.1 iJOIN Network Controller

The iNC is responsible of centralised control of the transport plane of the veNB. It is in charge of configuring, monitoring and driving the operation of the rest of the RAN and backhaul network entities. In particular, we adopt the widely followed approach in the industry by using the OpenFlow protocol [39] as SDN mechanism between the controller and the transport network entities iTNs as well as the iSCs. An extended OpenFlow controller is located at the iNC. It takes care of all the protocol interactions with the rest of the network entities. Each entity only needs to support standard OpenFlow using the so-called Southbound protocol interface (in iJOIN, this is the J3 interface). In addition to the pure forwarding computation intelligence, the iNC also hosts other functions. These functions support different management tasks such as mobility management (through the J4 interface to the MME), energy efficiency or load balancing (see D4.3 [14] for additional details).

The operation of the iNC can be summarised as follows. Once the functional split has been determined, a routing module in the iNC is responsible of computing all forwarding paths for each iSC-RANaaS connection. This computation takes into consideration the run time conditions of the network, which are obtained from continuous measurements and monitoring conducted in the backhaul. It further interacts with the energy efficiency, congestion control and mobility modules, such that a consistent and stable status is always maintained in the network. Note that traffic differentiation might be needed at the transport level because different functional splits may be executed over the same backhaul network. Similarly, different RANaaS instances might be used, which can be located in different places.

The required OpenFlow extensions are aligned with the work being carried out at the Wireless & Mobile Working Group of the Open Networking Foundation (ONF) [40], the standardization body in charge of the OpenFlow specifications.

5.3.2 iJOIN Transport Node

The iTN is an aggregation and forwarding node used to connect the iSCs with the RANaaS instance and the EPC. There are different technologies, e.g., fibre, Ethernet, or mmWave radio (see Table 7-1), that can be applied. Not all the technologies provide the same characteristics with respect to bandwidth and delay. Hence, the iNC controls the iTNs through the J3 interface to dynamically configure the optimal paths. In this computation, both the state of the network and the requirements of the traffic are considered.

5.3.3 iJOIN Local Gateway

The iLGW is a logical entity that enables local breakout by providing connectivity with local IP services which might even include Internet access. In this way, some traffic flows can be locally anchored at an iLGW in order to offload the operator’s core network. This entity is conceptually similar to the 3GPP Local Gateway [41].

5.4 iJOIN Functional Architecture

The two key concepts of the iJOIN framework are RANaaS and joint RAN/backhaul optimisation in dense small cell deployments. This section discusses iJOIN’s functional architecture, defining the interaction of functional blocks in the iJOIN architecture. This kind of definition allows for identifying the interactions across work packages which are of particular interest but also require special care to avoid inconsistencies.

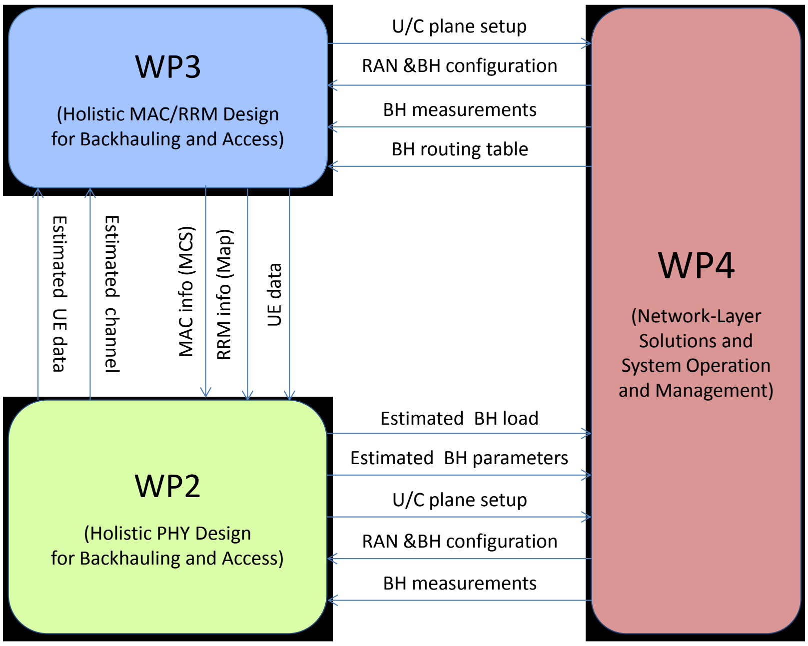

Figure 5-8: iJOIN WP Functional Interaction

Figure 5-8 shows an overview of the functional architecture from a project perspective. The interaction of WP2 and WP3 focuses on the exchange of estimated channel information such as SNR and user data after detection and decoding. WP3 mainly provides RRM and MAC information such as scheduling maps, link adaptation parameters, and block sizes. Furthermore, the user data needs to be provided. WP4 provides to both WPs information about the backhaul configuration and measurements such as expected throughput, routing information, and mobility information. WP2 and WP3 provide details about the U/C plane split to WP4 which needs this information for an appropriate routing through the backhaul network.

Figure 5-8 depicts a high-level overview of the functional architecture by focusing on interactions and information exchanged between the individual WPs, i.e., WP2 at PHY Layer, WP3 at MAC/RRM Layer and WP4 at Network Layer. The following subsections describe the detailed interactions between the specific modules and CTs within each WP. In addition, the deliverables D2.3 [8], D3.3 [11], and D4.3 [14] provide a detailed overview of CTs and related implementations from each individual WP perspective.

5.4.1 WP2: PHY layer

The PHY layer is responsible for the actual transmission of user messages over the wireless channel including generation of transmit signals and the detection at the receiver side. Within the project, a frequency reuse of one is envisioned leading to severe interference of signals across small cells. However, this interference can also be exploited by means of cooperative detection and transmission schemes. Thus, WP2 aims to develop novel PHY technologies that enable a flexible and distributed placement of PHY processing among the densely deployed network nodes. In particular, the development of novel distributed and cooperative PHY detection approaches in the uplink and transmission approaches in the downlink are in the scope. To this end, seven promising CTs have been proposed in D2.1 [3] and are discussed in detail in D2.2 [4] and D2.3 [8]. Table 5-2 lists these PHY CTs indicating their main function and the optimization area.

Table 5-2: iJOIN PHY layer Candidate Technologies

CT	Topic	Abbreviation	Function	Optimization area
2.1	In-Network-Processing	INP	UL Detection	RAN
2.2	Multi-Point Turbo Detection	MPTD	UL Detection	RAN
2.3	Joint Network-Channel Coding	JNCC	UL Detection	Joint RAN/BH
2.4	Sum-Rate and Energy-Efficiency Metrics of DL COMP with backhaul constraints	CoMP	DL CoMP	RAN
2.5	Partially Centralized Inter-Cell Interference Coordination	ICIC	DL CoMP	Joint RAN/BH
2.6	Data Compression over RoF	RoF	Functional Split/BH	BH
2.7	Millimeter wave backhauling	mmWave	for UL and DL	Joint RAN/BH

Figure 5-9 visualizes the interaction of the proposed WP2 CTs with other basic PHY layer functions, with MAC layer and with network layer functionality. In principle, the WP2 CTs can be clustered in UL and DL approaches for the iJOIN architecture.

CT2.1 and CT2.2 implement joint multi-user detection schemes in the uplink in order to cooperatively estimate the messages of several users in the vicinity of iSCs within the veNB. In particular, CT2.2 considers centralized as well as local turbo detection with only limited or no exchange over the J2 interface. In contrast, CT2.1 requires high profile J2 interfaces for cooperative detection with only a minimum of J1 communication. The joint optimisation of access and backhaul transmission for the UL is considered by CT2.3 and CT2.7 by developing novel joint forward error correction coding techniques.

CT2.3 develops a novel joint forward error correction coding approaches for access and backhaul transmission. The network coding approach is designed to allow for improving access performance while reducing the backhaul data rate when two pairs of smalls cells detection two pairs of users jointly. In contrast, CT2.7 jointly designs the access and backhaul transmission taking mmWave backhaul techniques into account.

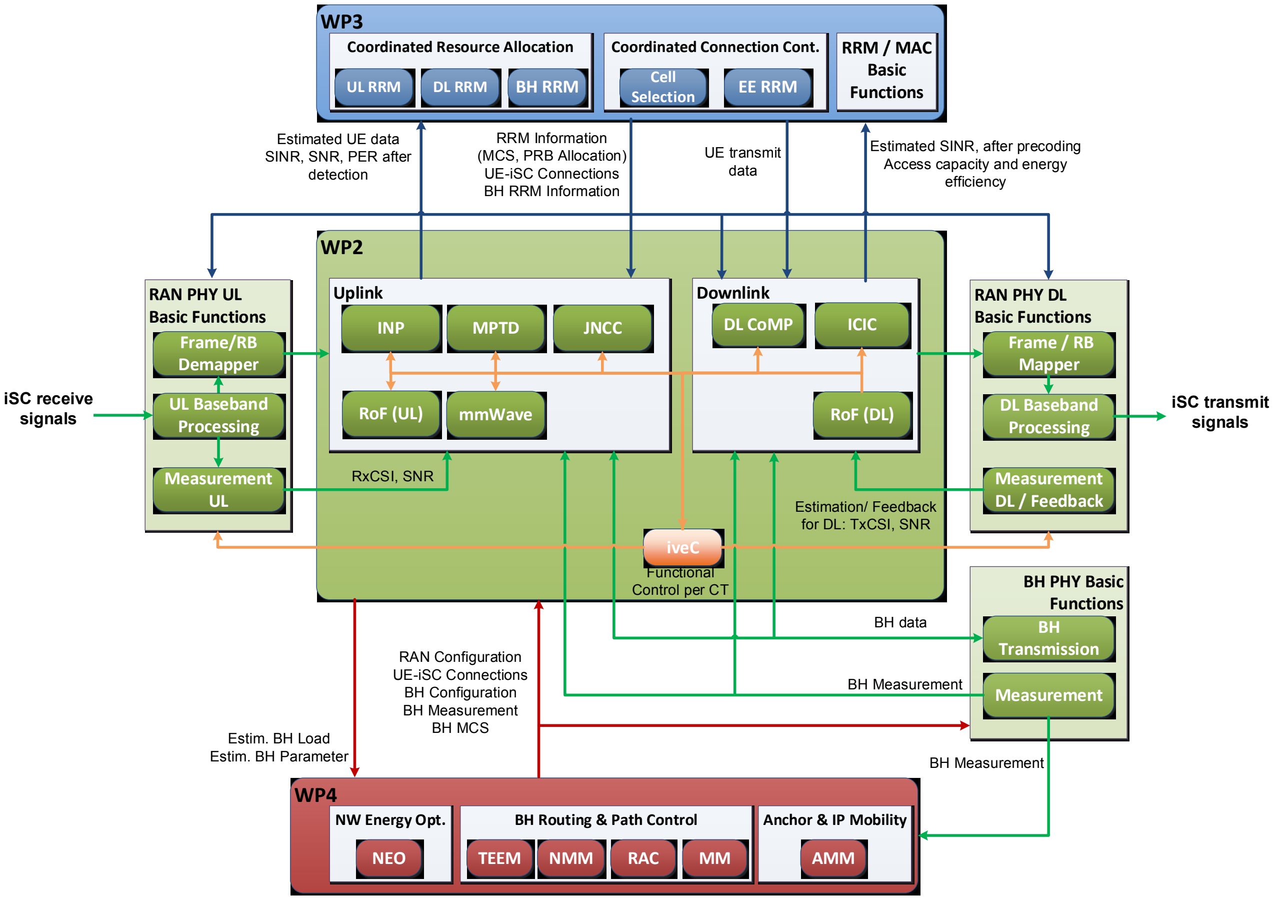

Figure 5-9: Functional Architecture of PHY processing

For the downlink again centralised as well as distributed approaches for joint multi-point transmission are considered in order to adapt to different BH technologies. CT2.4 investigates centralised precoding techniques and optimises the number of cooperating iSCs to maximize the area throughput taking realistic complexity constraints and BH constraints into account. For the case of distributed precoding, CT2.5 develops a new precoding scheme which is robust to channel imperfections. The forwarding of signals between network nodes of one veNB is a central property of the iJOIN architecture. In order to support the different UL and DL approaches, CT2.6 investigates different forwarding techniques by means of different functional split approaches.

These novel iJOIN specific approaches work together with basic PHY functions such as baseband processing, mapping and demapping of resource blocks and channel estimation which are not part of WP2 investigations. Thus, standard implementations are assumed as supporting blocks for the UL and the DL. Nevertheless, the iveC function controls also some of these basic blocks, e.g. the location of channel estimation depends on the actual functional split. Thus, the measurement needs to be executed in the iSCs or in the RANaaS instance. Based on the de-mapper output signal and the channel state information on the UL, alternative CTs for UL detection are considered.

5.4.2 WP3: MAC layer

WP3 is concerned with MAC, RRM and RRC techniques in RAN and in the backhaul network, including radio resource allocation, QoS scheduling, interference coordination, and connection control e.g. for handover or load balancing. In D3.1 [9], D3.2 [10], and D3.3 [11], joint optimisation techniques have been developed which use information exchange between network domains (RAN, BH, RANaaS platform) and functional blocks considered in WP3 as well as in WP2 and WP4.

The WP3 CTs are listed in Table 5-3 and are classified according to their specific functionalities and functional split configurations. In particular, CTs 3.2 and 3.3 can be characterised as SON functionalities, which enable coordinated connection control. They adapt the system parameters to changes in the cellular network, e.g., due to the network load, energy constraints, and mobility.

The other CTs are used in the centralised resource allocation framework. In particular, CT 3.1 enables optimised 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 investigates the iJOIN utilisation and energy efficiency metrics, which enable to assess the improvements of the proposed CTs. Hence, this classification it is not applicable to CT3.6.

Table 5-3: iJOIN RRM/MAC layer Candidate Technologies

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

Figure 5-10 illustrates the functional interaction of WP3 CTs (the blue box) as well as the exchange of information towards WP4 (in red) and WP2 (in green). From WP2, WP3 CTs 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 regarding the radio access and the backhaul, e.g., scheduling maps and link adaptation parameters. WP2 functions forward to WP3 functions estimated radio and backhaul channel information such as SNR and user data after detection and decoding.

The exchange of information between WP3 and WP4 functions can be divided across two iJOIN logical entities: the iNC and the iTN. WP4 functions provide WP3 functions the 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 5-10 also the interaction of WP3 CTs with the iveC, which is the logical entity that adapts the functional split configuration according to system objectives and constraints (see Section 5.2.1.2).

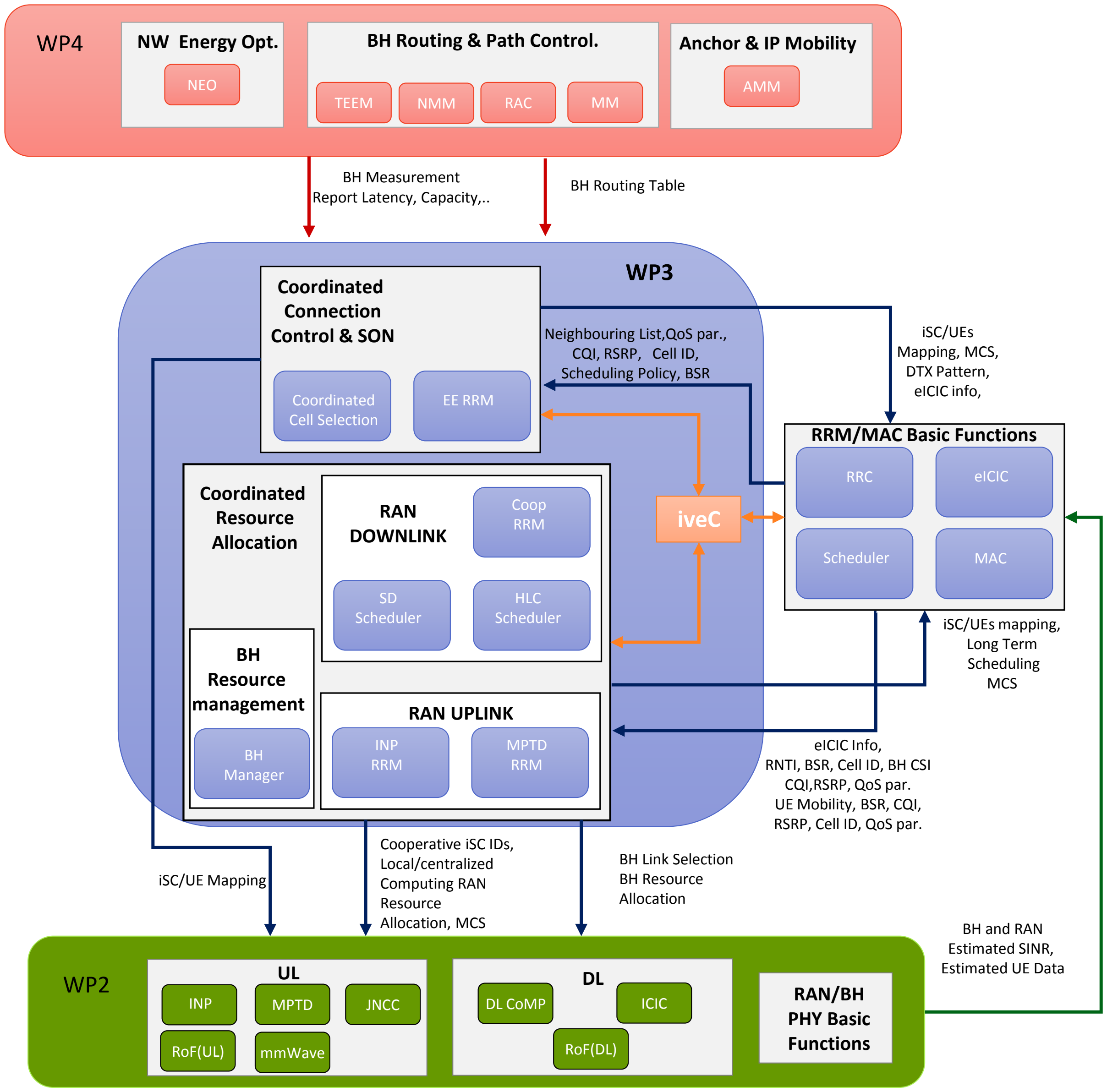

Figure 5-10: Functional Architecture of MAC/RRM

5.4.3 WP4: Network layer

WP4 is responsible for defining the mechanisms required to allow for an efficient operation of the network layer in terms of cost, energy consumption, mobility support, or latency. These activities have resulted in the definition of a set of Candidate Technologies (CTs) that were introduced in D4.1 [12] and then described in detail in D4.2 [13].

The network layer, in the context of WP4, is constituted by the set of elements that provide the connection of iSCs with the rest of the network, RANaaS instance, and S/P-GW with a certain QoS. In traditional networks, the base station or remote radio head and the rest of the network are connected either through the standard S1/X2 interfaces or the CPRI interface for C-RAN. In the context of iJOIN, it is possible to have different kinds of information and QoS associated to different functional splits, which adds a new level of complexity and requires new solutions for the support of functionalities such as mobility management, load balancing, routing, and congestion control.

The WP4 CTs are listed in Table 5-4. CT4.1 proposes a novel anchor selection and mobility management mechanism, having as a goal to select the optimal mobility anchors for the user based on several aspects such as backhaul status, support for local breakout, distributed anchoring, or user terminal and application specific aspects. CT 4.2 proposes a novel energy-efficient mechanism that considers not only the access but also the backhaul networks. CT4.3 investigates the fundamental problems of RANaaS positioning and dimensioning, i.e., answering the questions where to place RANaaS instances and how many resources to provide. This depends on the transport network structure, number of connected iSCs, data processing capabilities and side-constraints on latency and throughput, which result from the chosen functional split. CT4.4 proposes a novel congestion control mechanism. It allows for the reuse of the same transport infrastructure to support information flows that have very different latency and capacity requirements as well as operational time scales. The mechanism is based on the use of a virtual queue with several configurable thresholds, as well as the use of a selective random packet dropping mechanism and a packet pacer required to solve potential incast issues. CT 4.5 proposes a novel load-balancing mechanism that considers not only the user perspective but also the network perspective.

Table 5-4: iJOIN Network Layer Candidate Technologies (CTs)

CT	Topic	Function
4.1	Distributed IP Anchoring and Mobility Management	IP anchor selection and local breakout
4.2	Network Wide Energy Optimisation	Joint RAN/BH Energy Optimisation
4.3	Joint Path Management and Topology Control	RANaaS dimensioning and placement
4.4	Routing and Congestion Control Mechanisms	BH congestion control
4.5	Network Wide Scheduling and Load Balancing	BH load balancing

Figure 5-11 shows the functional interactions of WP4 CTs (the red box) among each other as well as with WP3 CTs (in blue), WP2 CTs (in green) and with the 3GPP MME entity. There is only limited interaction with WP2 CTs, mainly regarding to BH measurements. The exchange of information between WP4 and WP3 functions, as introduced before, can be split into two main parts: exchange of measurements and the configuration of the user and control plane through the iveC. WP4 needs to further interact with the MME to get triggers about mobility events, as well as to ensure a consistent state in the network in terms of mobility management. More details about the interactions of WP4 can be found in D4.3 [14].

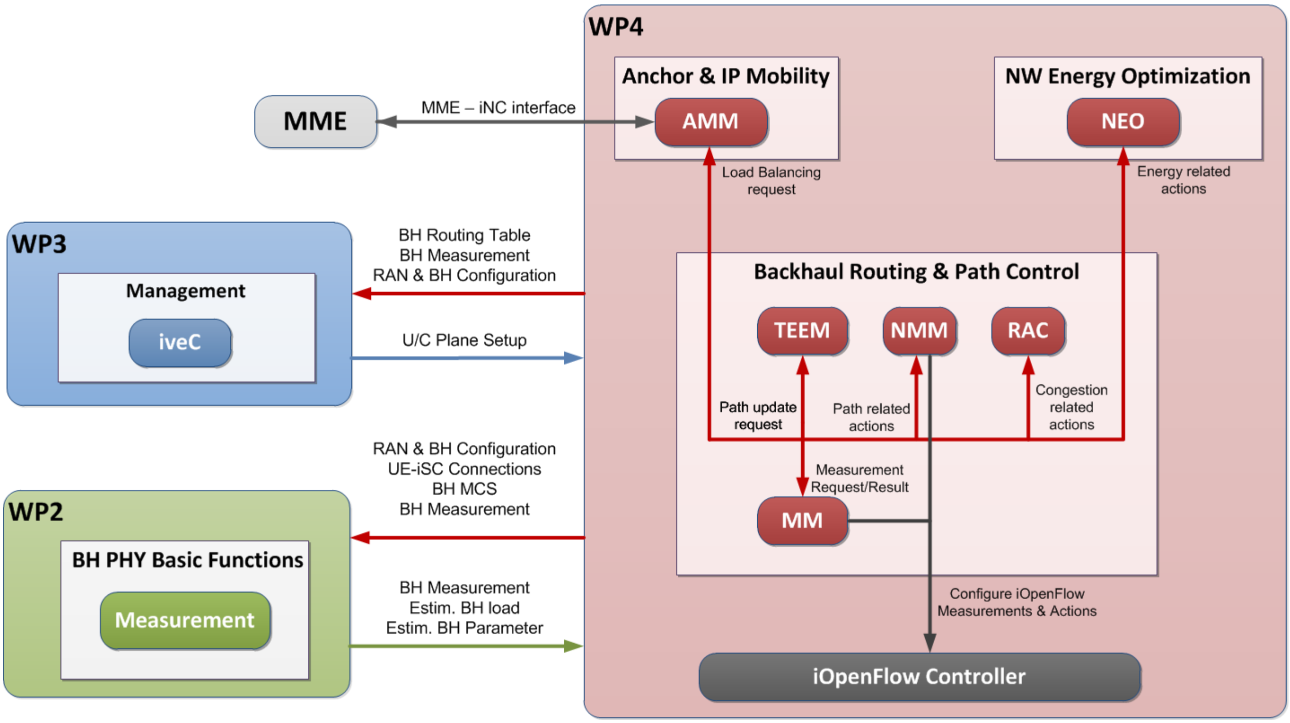

Figure 5-11: Functional Architecture of Network Layer functionality

5.5 iJOIN Physical Architecture

A generic backhaul scenario considered by iJOIN is shown in Figure 5-12: and was previously presented in D4.2 [13]. We consider both wired and wireless links for the interconnection of small cells with the metro network. Different deployment options are possible depending on the scenario. For example, iSCs might be connected via wireless links to one iTN and from there with fibre or cable based Ethernet to the aggregation network. Alternatively, a multi-hop wireless network might be used to provide connectivity to the aggregation or metro network. In order to be able to perform a characterisation of the latency and bandwidth of a given deployment scenario, it is important to study the latency and bandwidth characteristics of several access technologies that can be potentially used in the backhaul network.

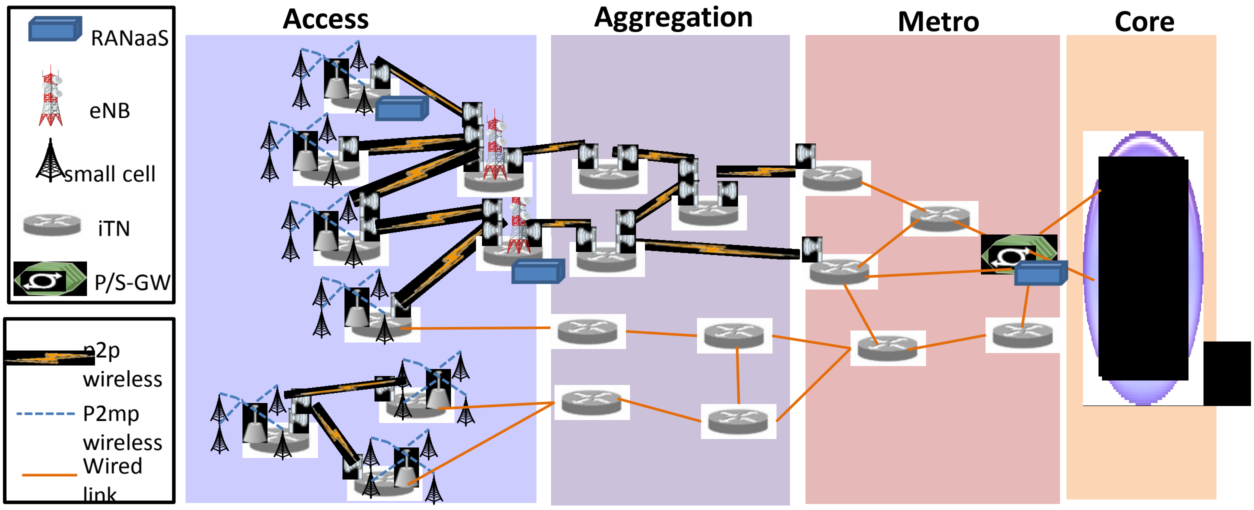

Figure 5-12: iJOIN generic backhaul scenario

The deliverable D4.3 [14] provides some examples of transport networks that, based on the discussions within the iJOIN consortium, may be considered as good examples of current deployment policies. Such examples are the starting point for defining a general framework for the characterisation of the RANaaS deployment options. Hence, the framework is based on potential deployment scenarios taking into consideration operator practices and forecasts. For each of the scenarios, the general framework is intended to provide an indication of likely values of a set of parameters, e.g., bandwidth, latency, topology and BH technology used. Based on the parameters and the framework described in D4.3 [14], we propose the transport network represented in Figure 5-13.

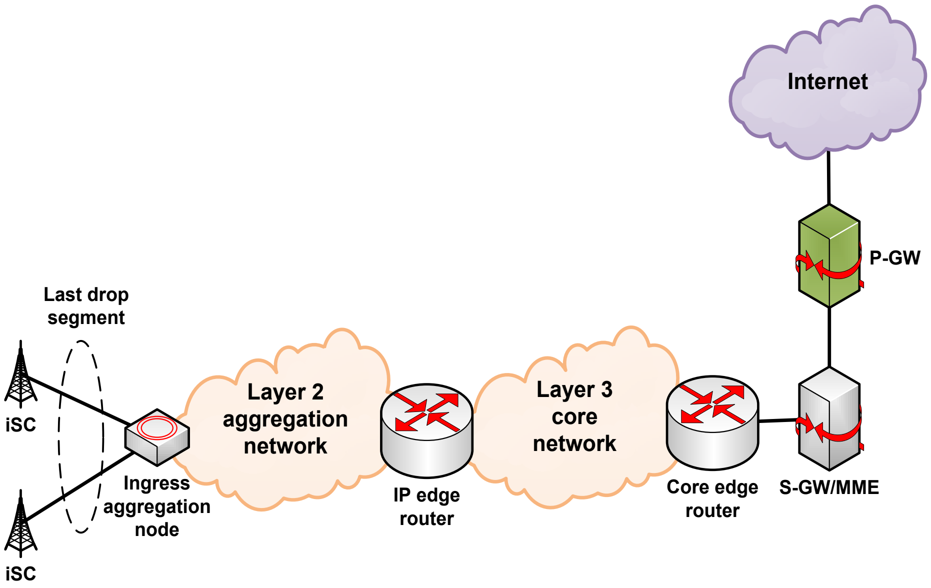

Figure 5-13: Proposed transport network

The proposed transport network is composed of three segments: the last drop segment, the metro aggregation network, and the core transport network. It is important to highlight that during the design of the network, the capacity estimation between the iSCs and the RANaaS infrastructure should take into account the limitations not only of the last drop segment, but also of the aggregation and core transport networks. Similarly, the latency estimation should also take into account not only transmission and processing delays at the edge of the network, but also the contribution of the intermediate nodes of the transport network.

This classic configuration is expected to change in the future as depicted in Figure 5-14. The development of low cost long reach optical transport solutions is expected to allow for a radical transformation of this existing framework into a different one with only two segments: the last drop segment, which would be long reach, and the core network.

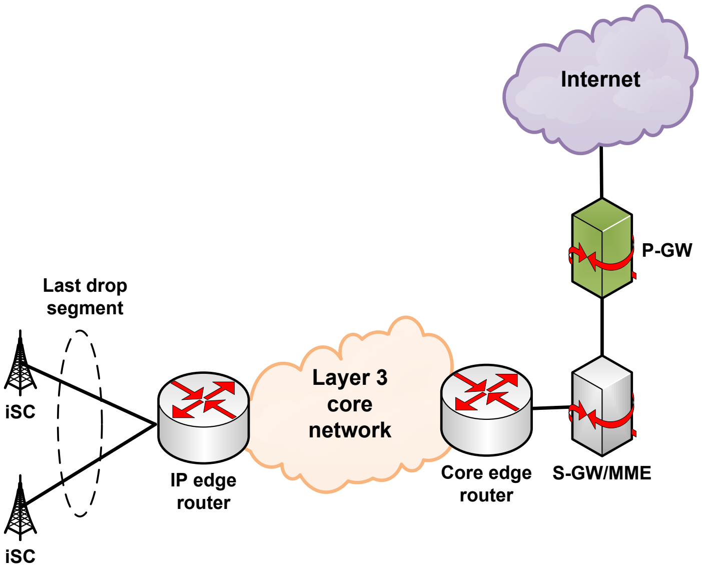

Figure 5-14: Evolved transport network

We refer to D4.3 [14] for further details on the proposed transport networks. For each Common Scenario defined in D5.2 [15], we provide in the following further details regarding typical physical deployments.

5.5.1 Common Scenario 1: Stadium

This scenario is assumed to be characterized by the deployment of a high number of radiating elements in a small area such that a very high traffic demand can be observed. The support of the very high traffic demands justifies the deployment of a dedicated infrastructure, both from a technical and an economic viewpoint. However, this scenario presents some unique characteristics. The premises proprietor deploys and owns the passive transport infrastructure. Hence, this passive infrastructure should be shared by different operators and used to support different radio access technologies. For this reason, a solution based in distributed antenna systems and radio over fibre is usually implemented in this kind of scenario.

For this scenario, we consider a last drop capacity and latency of 1-10 Gbit/s and 5-100 µs respectively. The last drop employs a 10 GbE dark fibre transport technology following a point-to-point topology. The last drop is directly connected to the core network allowing for a potential deployment of local breakout for Internet traffic. Figure 5-15: depicts a possible physical deployment for this scenario. We refer to Section 4.3 of D4.3 [14] for further details on this scenario.

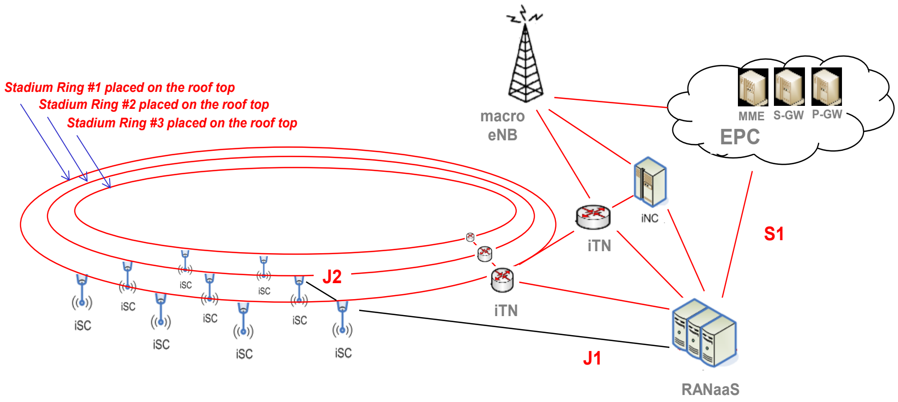

Figure 5-15: Stadium – Physical deployment example

5.5.2 Common Scenario 2: Square

For this scenario, the proposed RANaaS deployment is based on the standard network with the three segments indicated above: last drop, aggregation network and core network. The last drop will be based on the use of wireless backhaul solutions, e.g. microwave or millimeter wave, encompassing point-to-point and daisy chain topologies. The last drop segment length is 100-200m and a ring/mesh is assumed for the aggregation network. Aggregation network capacity is considered to be the lowest value of the typical network capacity (e.g., 10 Gbit/s). Figure 5-16: depicts a possible physical deployment for this scenario. We refer to Section 4.4 of D4.3 [14] for further details on this scenario.

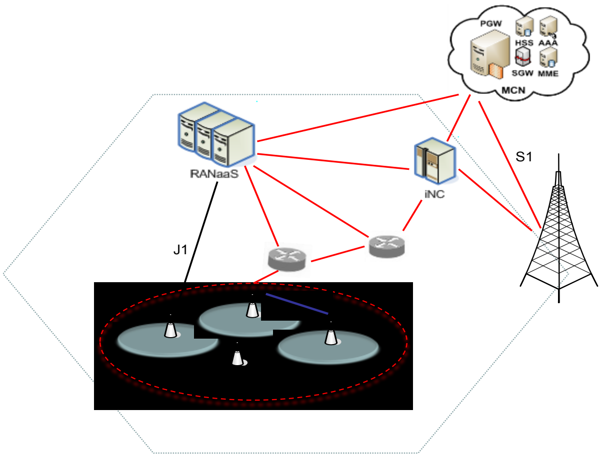

Figure 5-16: Square - Physical deployment example

5.5.3 Common Scenario 3: Wide-area continuous coverage

For this scenario, an evolved transport architecture is proposed that corresponds to the expected long term evolution indicated in Figure 5-14. This scenario envisages a long reach last drop, with a distance from the iSC to the core network of up to 10 km. The average distance is estimated to be 1.5 km and point-to-point and daisy chain topologies are employed. The last drop capacity is expected to be 10 Gbit/s per link. Moreover, the last segment offers Multiprotocol Label Switching (MPLS) functionalities (i.e., the iSC incorporates Provider Edge router functions). Wide-area scenario does not envisage any aggregation network, hence, the last drop is directly connected to the core network. Figure 5-17: depicts a possible physical deployment for this scenario. We refer to Section 4.5 of D4.3 [14] for further details on this scenario.

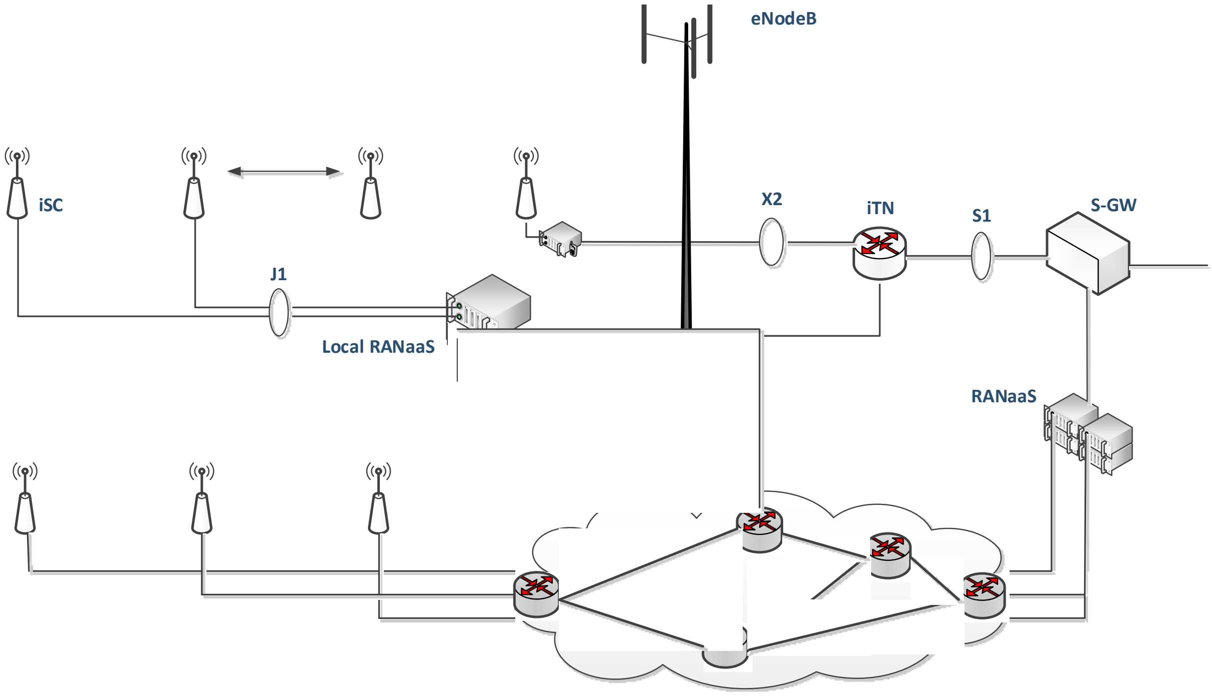

Figure 5-17: Wide Area - Physical deployment example

5.5.4 Common Scenario 4: Shopping Mall / Airport

In order to differentiate the shopping mall and Airport scenario from the Stadium scenario, it is proposed to assume that the solution has a lower performance for the transport infrastructure. This is in line with the fact that the traffic density that can be expected in this scenario would be lower than for Stadium scenario. The transport infrastructure may be provided by the premises owner, but it should not necessarily be dedicated to the mobile network, i.e., it may be employed for other uses. For instance, infrastructure should not necessarily be shared by operators and there may be limitations in the processing infrastructure that may be deployed in the premises.

Based on these assumptions, we envision 1 GbE last drop segments offering a capacity and latency of 1 Gbit/s and 5-100 µs respectively. The last drop is directly connected to the core network through point-to-point and daisy chain topologies enabling a potential deployment of local breakout for Internet traffic. Figure 5-18: depicts a possible physical deployment for this scenario. We refer to Section 4.6 of D4.3 [14] for further details on this scenario.

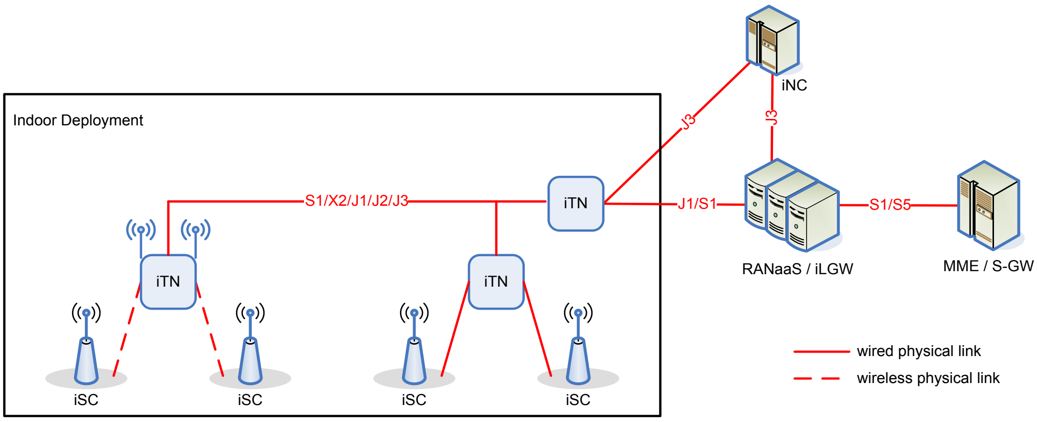

Figure 5-18: Shopping Mall / Airport: Physical deployment example

6 iJOIN Functional Split Concept

6.1 Motivation

Very dense and low-power small cell networks are a promising way to allow for handling future data rate demands. Firstly, the distance between the iSCs and UEs decreases and thus path-loss is reduced significantly. Hence, maintaining the same QoS, higher MCS rates can be applied increasing the data rate significantly. Secondly, small cells allow for the use of aggressive frequency reuse factors compared to large cells, which means that more bandwidth becomes available in each cell. Thirdly, small cell deployments cause each user to be in the vicinity of multiple small cells which allows for collaboration and cooperation between neighbouring cells.

Without applying sophisticated processing techniques where neighbouring cells cooperate, the deployment of small cells leads to strong inter-cell interference (ICI) in both uplink and downlink, which limits the performance gains. In contrast, centralised processing permits the implementation of efficient radio resource management (RRM) algorithms which allow for radio resource coordination across multiple cells. It also allows for optimisation of the radio access performance at signal-level, e.g. by joint multi-cell processing and inter-cell interference coordination (ICIC). Performance evaluations in [38] reports a gain in uplink area throughput of 20-50% near the cell-centre and up to 100% near the cell-edge.

Centralised-RAN (C-RAN) is one possible way to centralise computational resources, and it has recently attracted a great deal of attention. In C-RAN, multiple sites are connected to a central data centre where all the baseband processing is performed. However, C-RAN has very high requirements on the BH link between the iSCs and the central processing unit in terms of capacity and latency, such that only optical fibre is a suitable choice for deployment. For example, to forward the IQ (in-phase and quadrature) signals between iSC and RANaaS platform in case of B=20MHz bandwidth and 2 antennas at the iSC, the rate on the CPRI-based BH link equals 2.46 Gbit/s [4]. In addition, CPRI defines a BH latency of 5 µs to allow for coherent processing of signals. This contradicts with the expectation that future small cell networks will use a heterogeneous set of backhaul technologies, depending on factors such as cost, regulations, and availability (see Section 8.2.5).

Centralisation of mobile networks is an attractive way to lower the operational and capital expenditures of mobile networks [29]. The main source for lower operational expenditures (OPEX) are higher energy efficiency resulting in lower energy-costs, lower costs for site rental, and lower costs for management and administration of mobile networks. Reduced capital expenditures (CAPEX) may be mainly attributed to lower equipment costs. However, reliable information on OPEX and CAPEX structure of an existing mobile network is not available. In [37], the authors perform a comprehensive CAPEX analysis taking into account equipment costs for data centres, backhaul deployment, and radio access equipment. Based on the analysis in [37], this document provides a CAPEX comparison taking also into account gains through centralised processing (CoMP). It shows that CAPEX can be reduced by 10-25% (see Section 8.2.5) if iJOIN's key concepts are applied.

Thus, in addition to the extreme cases of a fully centralised and a fully decentralised architecture, more flexible concepts are required. These concepts enable a flexible function placement in the network depending on the deployed RAN and BH network, the current communication needs and available processing resources. For example, some lower PHY processing may already be applied in the iSC to reduce the BH rate, whereas higher layer PHY functions are executed in the RANaaS platform to benefit from centralisation gains. Furthermore, by introducing communication links between the iSCs via the J2 interface distributed as well as cooperative approaches become feasible. In principle, this allows for performance results comparable to centralised processing while limiting the computational needs in the RANaaS platform and the BH requirements. Thus, many more processing options for densely deployed small-cell networks can be realised using a flexible functional split. Table 6-1 provides a summary of principle processing approaches.

Two of the main concepts of iJOIN, the RANaaS implementation on commodity hardware and a joint design of access and backhaul enable and complement a flexible functional split. The utilisation of easily reconfigurable commodity hardware enables a system architecture where the functional split can be changed flexibly in time and space, selecting the most preferable option e.g. depending on time of day, availability of backhaul, deployment density, or traffic demand. The current two extreme architecture force operators to choose between an expensive FH network and waiving of centralised processing gains. The joint design of access and backhaul utilises this flexibility to optimise the network according to the current split. Of course, it needs to be carefully evaluated, which functionalities can be implemented on virtualised commodity hardware and how the functional split is decided. A detailed discussion on that can be found in Section 5 of D5.2 [15].

Table 6-1: Principle processing approaches for distributed systems

Processing Approach	Description	J1	J2	Complexity	Performance
Local	Each iSC processes locally without any cooperation	Low rate, medium latency	-	iSC with full complexity; RANaaS only for coordination	Baseline performance
Centralized	Fully centralised processing	Very high rate, ultra-low latency	-	very low iSC complexity; high complex RANaaS	Under ideal assumptions best performance
Distributed	Fully distributed processing among iSCs	Low rate, medium latency	Very high rate, ultra-low latency	Large complexity iSCs; RANaaS only for coordination	Depend on actual BH rate and latency
Cooperative	Processing among iSCs combined with RANaaS processing	Low rate, medium latency	High rate, Low latency	Scalable	Depend on actual BH rate and latency

6.2 Functional split Concept

The currently considered fully centralised and fully decentralised architectures are two extreme concepts, both with disadvantages. A decentralised network requires relatively low BH capacity but does not allow for centralised processing which may significantly improve network efficiency. A centralised radio access network enables joint processing techniques such as multi-user detection and coordinated multi-point transmission, yet it implies much more stringent requirements on the BH, i.e. BH links usually require several Gbps and latency in the order of a few microseconds. Furthermore, BH links are not compatible with existing transport network technologies, which increases CAPEX significantly.

Hence, there is the need for technologies which offer a centralisation gain while relaxing the requirements on the transport network. This can be achieved by splitting the RAN functionality into two parts, one executed locally at the iSC and one executed at the central processor. Depending on the chosen split, the BH requirements are reduced and a different degree of centralisation gain is achieved. There are many factors that determine the required data rate for each split and there are numerous options to split the RAN processing. In iJOIN, four main functional splits have been considered (depicted in Figure 6-1), with additional sub-splits being discussed in D2.3 [8] and D3.3 [11]. The matching between functional split options and iJOIN CSs is detailed in D4.3 [14].

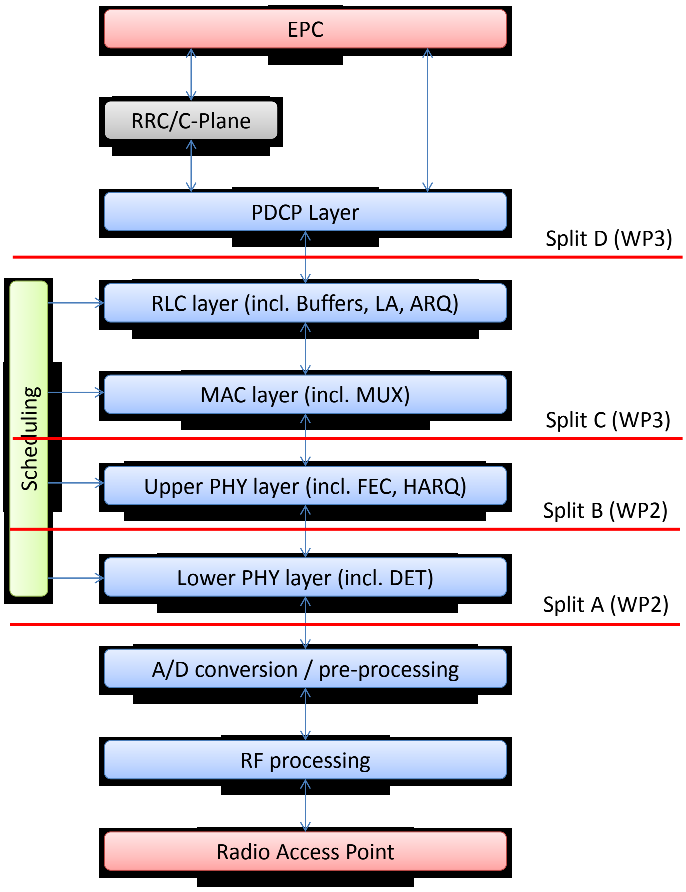

Figure 6-1: Main functional split options considered in iJOIN

The split options are analysed in the following with a special focus on the uplink as it gives more options for joint processing. When not specifically noted the extensions to the downlink are straight forward.

Split A

This option corresponds to a full centralisation known from C-RAN. CPRI is a commonly used standard for this option. In the uplink, the received signals are down-converted to baseband and converted to the digital domain. As the complete baseband signal is forwarded, the required FH rate is quite static and it is determined by the system bandwidth, the number of receive antennas per sector, number of carriers, and number of sectors per site. Furthermore, the required bit-rate scales linearly with the bit-resolution of the analog-to-digital (A/D) conversion (and vice-versa), which is usually chosen to be 15 bits per symbol because of the high peak-to-average power ratio of the time domain signal and to ensure precise channel measurements. The main drawback of Split A is the independence of FH data rate and actual user traffic, i.e. even BSs that currently serve no users will require the full FH capacity. Also it is very challenging to implement all baseband processing on commodity hardware, which is why current C-RAN solutions use less flexible dedicated hardware such as ASICs and DSPs in the central entity.

Obviously, this split has no restrictions regarding the type of centralised processing that can be performed. Furthermore, apart from digital filtering and the BH protocol, no local processing is required. On the other hand, time and frequency synchronisation is performed centrally, which imposes strong latency requirements. CPRI defines a round-trip time (RTT) of 5 μs plus propagation delay on the FH, which sums up to a few hundred microseconds at typical distances of a few tens of kilometres.

An example of functional split A on the uplink is the Multi-Point Turbo Detection (MPTD) approach developed in CT2.2 together with the appropriate RRM designed in CT3.7. For this CT, the receive signals for cell-edge users are forwarded to the RANaaS to allow for an iterative joint turbo detection of two users served by two iSCs. The system level evaluations indicate a gain in area throughput of 40% to 60% compared to local detection. For the downlink, CT2.4 implements centralised joint transmission which achieves gains in area throughput under ideal conditions of 40% where backhaul delay, precoding matrix calculation delay, and CSI feedback delay are considered.

Split B

In the case of Split B, the received uplink signal is pre-processed, e.g. the CP is removed, the signal is transformed to the frequency domain, equalised, and the logical channels are de-mapped and detected. The main benefit of this split is that non-utilised resources such as guard carriers and unoccupied RBs are not forwarded to the central processor. In the case that synchronisation and equalisation is performed at the iSC, even reference symbols and synchronisation signals can be omitted. By contrast to split A, this split allows for exploiting the statistical multiplexing gain in the backhaul on the basis of occupied physical resources. Hence, the required BH data rate scales with the actual data traffic and channel quality. Furthermore, as analysed by [32] the frequency domain signals also require a lower bit resolution, i.e. the amount of quantisation bits can be reduced to 7-9 bits per symbol (depending on uplink and downlink). The output of the equalisation and detection stage in the uplink are soft-information values which are maximum 3 bit per coded bit, i.e. 6 bit for QPSK, 12 bit for 16-QAM and 18 bit for 64-QAM. The input of the modulator in downlink is also tied to the modulation scheme as the coded bits per symbol need to be forwarded, i.e. 2 bit for QPSK, 4 bit for 16-QAM and 6 bit for 64-QAM. Furthermore, MIMO processing is performed so that the number of antennas is mapped to a number of spatial layers (or vice versa). Link-adaptation and MIMO scheme selection are performed depending on the channel quality. Hence, the required BH rate depends on the signal-to-interference-and-noise-ratio (SINR) and channel rank experienced by each user terminal. This reduces the required BH rates significantly for lower-rank channels.

For split option B, different degrees of cooperation among iSCs can be distinguished introducing different kinds of limitation. Without cooperation among iSCs via J2 links, only localised equalisation would be possible to provide estimates for the user messages. Thus, only by forwarding these estimates to the RANaaS some minor centralisation gains are possible by means of FEC decoding. In WP6 [33] [17], it was analysed that an implementation of turbo decoders on commodity hardware seems feasible. At the same time, a joint decoding scheme for access and backhaul was developed by CT2.7 in D2.2 and [34]. It makes use of split B so that this option promises to achieve significantly lower BH requirements while embracing the two basic concepts of iJOIN. By the application of this joint RAN/BH coding approach gains of 110% in area throughput are achievable.

If, on the other hand, cooperation among iSCs is possible, joint multi-user detection approaches or joint transmission approaches can be implemented in principle leading to considerable data exchange over the J2 interfaces. Candidates are CT2.1 where cooperative MUD is implemented by the In-Network-Processing approach and gains of approximately 50%-70% compared to local detection are reported in D2.3 [8]. In case of joint DL transmission, functional split B is considered by distributed precoding in CT2.5.

A limitation for this split in terms of latency originates from the Hybrid Automated Repeat Request (HARQ) process. It requires decoding acknowledgements within 4ms which is a tight requirement, but can be addressed using opportunistic HARQ as investigated in CT3.4 [11].

Split C

In this split, all PHY layer processing is performed within the iSCs either locally or distributed. While this option seems to offer a much cleaner split between L1 and L2, it removes the option for any centralised joint PHY processing such as multi-user detection and CoMP schemes, which promise large benefits for interference mitigation. Hence, the main centralisation gain follows from joint scheduling, interference coordination, and path management techniques which may still have a significant impact on the network performance. The required FH capacity is closely tied to the actual user throughput determined by the user channel quality as the FEC redundancy is removed. The main bottleneck is represented by the FH latency which may not exceed a few 10ms. Otherwise, scheduling and link-adaptation may perform sub-optimal and imply performance degradation.

As functional split C terminates the PHY processing in the iSCs, distributed processing among iSCs is an option to realise gains close to centralised processing. Again, CT2.1 can be applied with forwarding after FEC decoding at the iSCs. This has also the advantage, that the decoding gain can be used to terminate the iterative detection approach reducing the total BH rate. As before, CT 2.5 can be applied in this scenario. In this scenario, the pre-coding is divided into a central and local part whereas the centralised part has a coordinating function. Using CT 2.5, gains of up to 25% can be achieved in the downlink. Furthermore, CT 3.9 also takes advantage of this functional split and implements a cooperative scheduling and interference control scheme. Using CT 3.9, gains of up to 20% can be achieved in downlink.

Split D

Split option D is characterized by a centralised RRC and PDCP protocol layer. This allows for centralised connection control and SON-like functions such as parameter tuning, e.g. for hand-overs and power control, load balancing as proposed in CT 3.2 [11], and efficient power management as proposed in CT 3.3 [11]. This corresponds to the “split bearer” option of the dual connectivity feature in 3GPP LTE. This functional split has 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. Consequently, functional split D allows for coordination of RRC functions on medium-term time scales, such as load balancing, cell selection, and hand-over. Nevertheless, significant gains can be expected, e.g. on energy efficiency, where a centralised coordination of cell DTX leads to up to 70% energy savings in a wide area scenario [11].

6.3 Constraints and Requirements

6.3.1 Data Rate Requirements

Numerical Example

To illustrate how the different splits impact the required FH data rates, Table 6-2 gives an example based on the uplink of a 10 MHz system. In the table, we also summarise the dependencies between the required data rate and basic system parameters.

Table 6-2: Exemplary FH rates and dependence of functional splits on LTE system parameters

Split	Quantiser resolution	Number of antennas	Num. spatial layers	Bandwidth (FFT size)	Occupied resources	Modulation scheme	Code rate	Required FH rate (for 10 MHz LTE, normal CP, 33kbit transport block, 64 QAM, 2 antennas, code-rate 0.85, one spatial layer, 25% FH overhead)
A	X	X		X				1.23 Gbps
B	X		X		X			155 Mbps
C			X		X	X	X	44 Mbps
D			X		X	X	X	44 Mbps

6.3.2 Aggregation Gains

The high costs for BH and FH networks require well dimensioned transport networks which are not over-dimensioned in order to minimize the deployment costs and not under-dimensioned to avoid outage and satisfy user expectations. In D2.3 [8], we derived the distribution of data rates of different functional splits based on traffic variation and channel quality variation. This distribution and the data rate requirements per functional split are used to derive the *aggregated* data rate requirements per functional split including the statistical multiplexing gain.

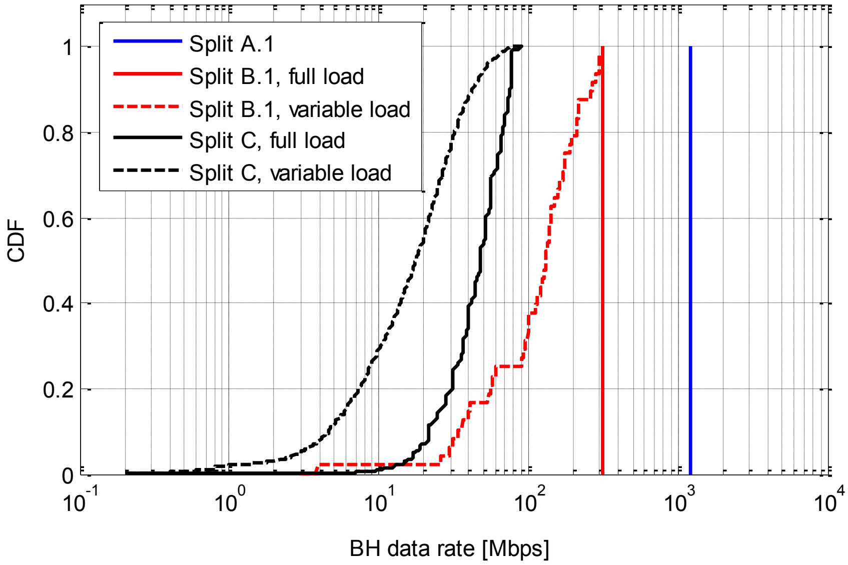

Figure 6-2: CDF of BH data rates with full and variable load

Figure 6-2 shows the distribution of required data rates for one base station and the three functional splits A, B, and C. These data rates may vary quantitatively in different scenarios and depend among others on the user density, small-cell density as well as propagation environment. However, the qualitative comparison and conclusions will remain unchanged. Evidently, the data rate of Split A.1 is constant and does not change with the actual traffic demand. Split B.1 varies with the actual traffic load as reflected by the curve for time-variant traffic. Split C additionally varies with the three modulation schemes, hence, its backhaul depends on the actual information rate. As split D is not fundamentally different from split C in terms of data rate requirements it is not depicted in Figure 6-2. These results show how higher functional split couple the BH traffic to the actual user traffic. The resulting variance in BH traffic can be exploited for aggregation gains.

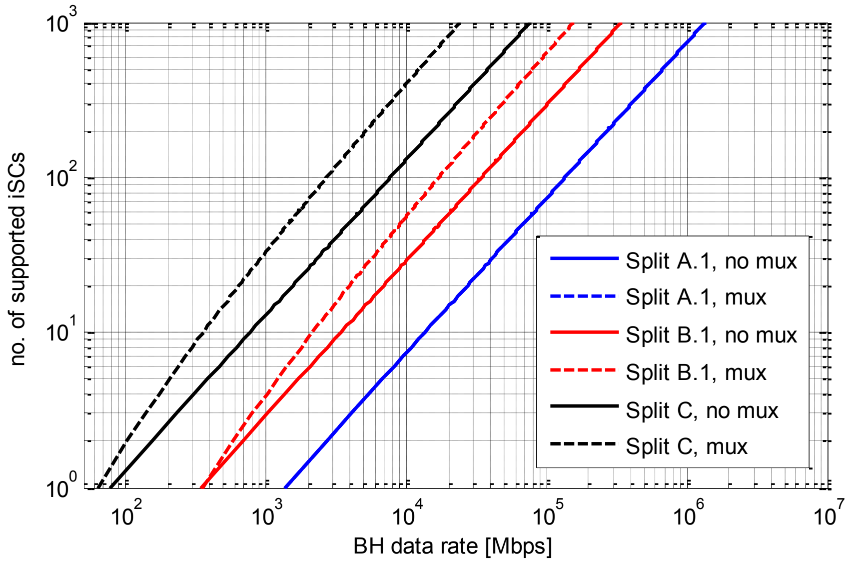

Figure 6-3: Number of supported iSCs per deployed aggregation capacity

Figure 6-3 shows the number of iSCs that are supported by a certain aggregation capacity. In case no multiplexing is considered, a number of iSCs can be supported if the deployed aggregation capacity corresponds to their combined maximum BH traffic. In contrast, when considering multiplexing, it is assumed that a number of iSCs can be supported if their combined BH traffic does not exceed 90% of the deployed aggregation capacity with a probability of 99%. This is common practice and yields a so-called statistical multiplexing gain. As can be seen, split A does not yield any gain as its BH traffic is static. Splits B and C do yield gains with Split C giving the highest gain of up to factor 3 according to its large variance. This clearly illustrates the benefit of higher splits as they not only reduce the required BH capacity in general but by varying according to the actual user traffic they additionally result in multiplexing gains.

6.3.3 Computational Constraints

One of the key challenges of RANaaS that has been tackled by iJOIN is the implementation of RAN functionality in software on commodity hardware, e.g. in deliverables D2.2 [4], D2.3 [8], and D5.2 [15]. Such an implementation requires a new way of designing and operating the RAN. So far, RAN functionality has been executed on dedicated hardware such as DSPs or ASICs. Those platforms are dimensioned to provide sufficient resources to cope with peak-traffic demands. They are further highly reliable and cost-efficient. On the other hand, those platforms do not permit sharing or virtualisation of resources. In contrast, software implementation on commodity hardware may be more flexible, allows for resource sharing, or virtualisation. However, it is usually less reliable and has poorer performance. Therefore, these implementations need to be "cloud-native" and must be designed for resilience. This cannot be achieved by porting existing implementations but requires more advanced concepts.

Beside general purpose processors (GPPs), also hybrid architectures may be deployed, i.e. GPPs that are complemented with network-processors for lower-layer processing, similar to GPUs in current computer architectures. Those network-processors may then be addressed through open interfaces (similar to OpenGL) in order to allow for flexibility and resource sharing within the Cloud-RAN. Additionally, the processing may be performed in virtual machines or in more light-weight environments such as containers. Furthermore, the processing depends on the processing granularity, i.e. whether it is performed on a per-user-terminal basis, a per-base-station basis, or a per-cluster basis. Per-user-terminal processing maximizes the scheduling degrees-of-freedom while implying the highest synchronisation overhead; per-base-station processing may simplify the merging of user-terminal data but may introduce significant synchronisation overhead for multi-cell processing; lastly, per-cluster-processing allows for joint processing but also decreases the degrees-of-freedom for scheduling of computational resources.

In a RANaaS system, the data processing requirements depend on many different factors. For example, if the transmission rate increases, more information bits need to be processed which, in turn, increases the computational complexity linearly. Additionally, if a communication link operates close to its Shannon capacity, even more receiver processing is required, which can be attributed to the need for additional turbo decoder iterations. As a result, the processing complexity (in fact) increases super-linearly as the system operates increasingly close to the maximum achievable capacity, and it depends on the instantaneous channel condition per user terminal (UE).

Therefore, in a manner similar to exploiting channel diversity in mobile networks (e.g., multi-user diversity in scheduling or spatial diversity in multiple antenna systems), *computational* diversity can also be exploited. Computational diversity refers to the large fluctuations in the data processing load due to changing channel conditions. Hence, if multiple users are served by a RANaaS instance, their diverse computational requirements may be used to improve the resource utilisation since the computational assets need to be provisioned according to the expected cumulative load of the users rather than the peak load of any given user.

When the data processing load exceeds the available computing resources, a *computational* outage occurs. From a user's perspective, there is no difference between a channel outage and a computational outage: In either case, the communications fail and another attempt must be made to transmit the packets. The model described in [35] accurately predicts the data processing resources required to perform uplink decoding in a multi-cell scenario. Assuming a 10MHz LTE channel and that each turbo-decoder iteration requires 1000 FLOPS per data bit, we can estimate the overall required data processing capabilities which is then compared with a reference setup, where server blades are equipped with four Intel Xeon 4870 (10 core processor) and 128GB RAM. Finally, we use the empirically determined distribution of computational load among RAN functions as discussed in [36].

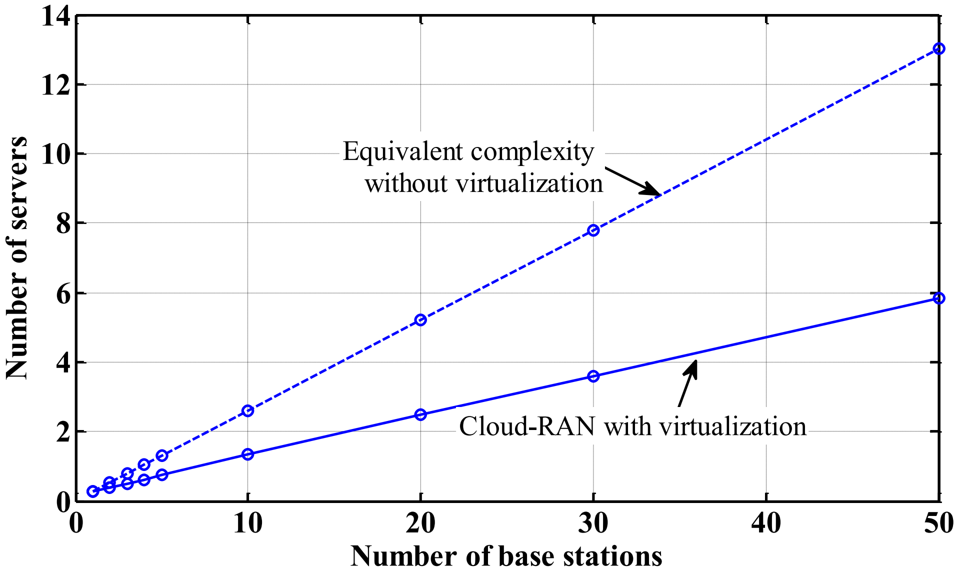

Figure 6-4: Exemplary results for computational diversity gain in a virtualised environment

Based on these assumptions, Figure 6-4 shows the computational resources required for LTE in a data centre using a shared pool of resources (i.e., RANaaS). In order to illustrate the computational diversity gains which can be exploited in such a system, we compare this with a non-virtualised case where each base station is serviced by its own dedicated computational resources. For both systems, cells are assumed to be fully loaded and the computational outage probability is set to 10%. When resources are shared, we see a reduction of about 50% in the required data processing resources.

In a RANaaS system, resources of the RAN and data-processing platform may be managed jointly, i.e., radio resource allocation must adapt not only to the RAN channel quality, user quality of service, and past performance, but it also must take the allocation's demand for computational resources into account. This is a predictive task, i.e., the system has to estimate the required computational complexity, estimate the available computational resources, and then adapt the RAN resource allocation accordingly. This joint optimisation of wireless and data processing resources has been addressed by CT 3.6 in D3.3, Section 4.6 [11]. CT 3.6 allows for operating the data centre at computational capacity, i.e., at the maximum system throughput with a required reliability using a given number of computational resources. Such a requirement may be particularly important during peak-traffic hours when a very large number of users connect to the mobile network and the required computational resources approach the system limit.

The previous examples described operational tasks. However, we also need to address dimensioning and positioning challenges. Dimensioning of RANaaS instances has been addressed in D5.2 [15] and positioning of RANaaS instances has been addressed in D4.3 [14]. Both tasks are important to make sure that data centres have sufficient resources to process incoming traffic but also to avoid over-provisioning. Furthermore, resilience is important in order to account for possible outages of data centres.

6.3.4 Timing Constraints

Different latency requirements need to be fulfilled by the backhaul for different functional split options. Since 3GPP defines many timers on all layers, these values define the maximum latency requirement needed per layer enabling a transparent functional split, i.e., without any specification changes.

In LTE, the PHY layer works with 1 ms subframe granularity. At the MAC layer, the HARQ timing is the most critical one. Once a subframe has been sent at subframe n for a given HARQ process, an acknowledgement (positive or negative) is expected at subframe $n+4$. 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 3 ms.

This is illustrated in Figure 6-5 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.

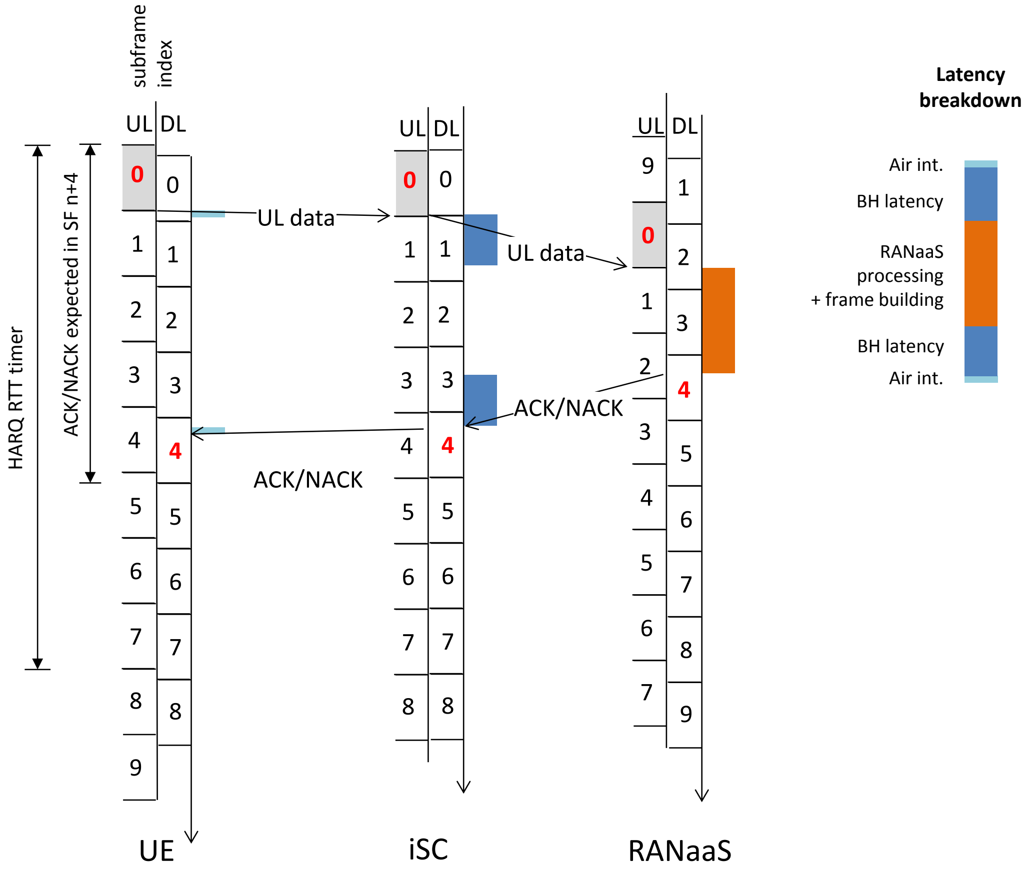

Figure 6-5: Break down of timing constraints for centralized HARQ in LTE FDD

Backward compatible solution exists to delay the retransmission but this will stall the HARQ process at the same time, reducing the throughput [43]. 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 [44] and in D5.2 [15].

The LTE MAC layer also 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 [11].

Table 6-3 shows the main specified timers and timing constraints per layer which may impact the functional split. Apart from the MAC timers, timers in RRC offer sufficient timing range to be configured for different backhaul latency enabling a functional split above the MAC layer without too many difficulties from implementation and system performance perspective. Nevertheless, these timers impact the overall system performance and require careful tuning to avoid performance degradations.

Table 6-3: 3GPP timing requirements

	Timer	Short description	Max Value
PHY	Subframe	Physical subframe length	1 ms (fix)
	Frame	Physical frame length	10 ms (fix)
MAC	HARQ UL indication	When an ACK/NACK indication is expected	FDD: SF+4 (3 ms) TDD: <= SF+7 (6 ms)
	HARQ RTT Timer	When an HARQ process is available	3 ms (fix)
RLC	t-PollRetransmit	For AM RLC, poll for re-transmission @tx side	500 ms
	t-Reordering	For UM/AM RLC, RLC PDU loss detection @rx side	200 ms
	t-StatusProhibit	Prohibit generation of a status report @rx side	500 ms
PDCP	discardTimer	Discard PDCP SDU / PDU if expiration or successful transmission	Infinity
RRC	TimeToTrigger	Time to trigger of a measurement report	5.12 s
	T300	RRCConnectionRequest	2 s
	T301	RRCConnectionReestablishmentRequest	2 s
	T304	RRCConnectionReconfiguration	2 s or 8 s
	T310	Detection of physical problem (successive out-of-sync from lower layers)	2 s
	T311	RRC connection reestablishment (E-UTRA or another RAT).	30 s

7 Joint RAN/BH

7.1 Motivation

In this section, we will demonstrate the need for a holistic design of RAN/BH functionalities by analysing conventional distributed and centralised architectures.

Specifically, the performance of a conventional RAN architecture is affected by the transport network capacity and latency characteristics. The BH capacity depends on the BH technology and the topology [5] used for connecting small cells. These parameters largely affect the amount of signalling information that can be exchanged to enable centralised coordination in lower layer functional splits [4]. Moreover, as we will discuss, BH technologies such as DSL or microwave can also introduce bottlenecks to the cell throughput.

The BH latency depends on the used BH technology as well as on the transport network topology. The requirement to exchange control information can be vital for the applicability of flexible centralisation because latency affects the CSI/CQI reliability. Therefore, interference management schemes implemented through centralised transmission and reception schemes and coordinated scheduling may not lead to notable gains when the transport network induces high latencies. In the same way, we have observed that a high backhaul latency increases handover preparation time, which may lead to an increasing handover failure rate [10].

Table 7-1: Backhaul Classification [13]

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	PtmP (LOS)	TDD	TDMA
3a	Sub-6 GHz Unlicensed or licensed		≤5 ms	≤500Mbps	PtP (NLoS)	TDD	
3b			≤10 ms	≤500Mbps (shared)	PtmP (NLoS)	TDD	TDMA
3c			≤5 ms	≤1 Gbit/s (per client)	PtmP (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	PtmP		TDM (DL) TDMA (UL)
5	xDSL		5-35 ms	10M – 100Mbps	PtP		
6	1 Gigabit Ethernet		≤200 µs	1Gbps	PtP		

Regarding the different technologies that can be employed in the backhaul, a thorough analysis is performed in D4.2 [13]. Table 7-1 summarises different backhaul technologies that are considered within the scope of iJOIN. The decision on the functional split and the implementation feasibility of the different candidate technologies will depend mainly on the latency and the throughput imposed by the technology available for the backhaul. In this regard, for wireless backhaul, the latency per hop and the achievable throughput depend on the range of frequencies employed, on availability of Line of Sight transmission and on the topology (PtP or PtmP). On the other hand, for fibre transmission, the main factor is the topology and the multiplexing technology. Figure 7-1 shows the relationship of these backhaul technologies and the preferred functional splits described in the previous chapter.

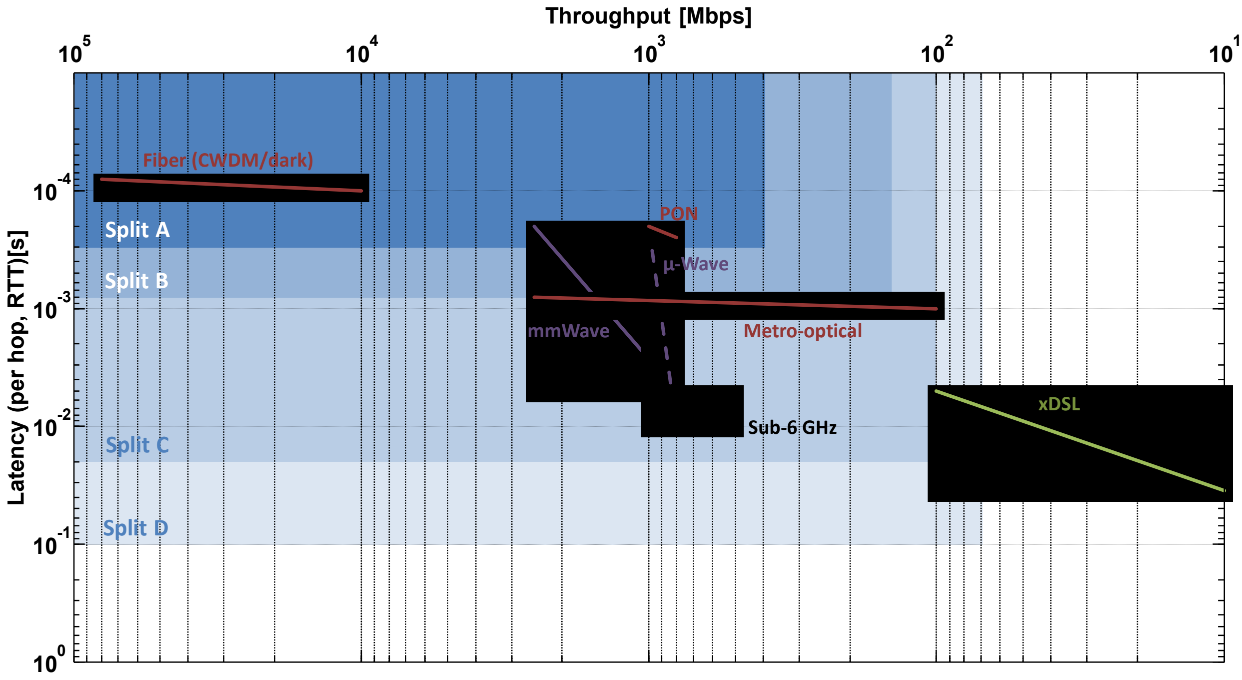

Figure 7-1: Relationship of backhaul technologies and functional splits

Backhaul Capacity

In current LTE technology, the association between a UE and a small cell is based on the strength of reference signal received power (RSRP). Due to the downlink power imbalance in heterogeneous networks, this solution limits the opportunities for macro cell offloading and the usage of the overall network resources. Recently, cell range expansion (CRE) jointly applied with ICIC has been proposed to improve fairness and capacity [7]. However, the cell load and the backhaul characteristics should be taken into account to optimise the cell association. Figure 7-2 shows the impact of different BH technologies on the maximum system area throughput when a classic RSRP scheme is used to manage the cell association. Results demonstrate how the BH capacity can be a bottleneck to the RAN performance and motivate a joint RAN/BH aware connection control mechanism [11].

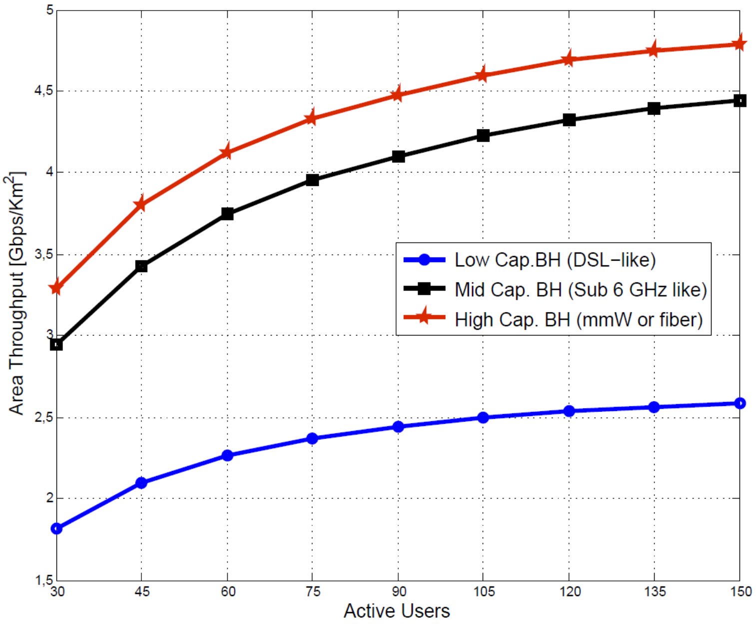

Figure 7-2: Maximum Area Throughput achieved by a basic RSRP Connection Control

Backhaul Latency

Figure 7-3 illustrates the area throughput performance in a C-RAN architecture as a function of the backhaul delay. In a conventional C-RAN architecture, central availability of CSI from all iSCs allows for coordinated scheduling. However, with increasing backhaul latency, CSI becomes more and more outdated, which affects the accuracy of the scheduling decision and the link adaptation. Especially the latter aspect is responsible for the large performance drop, obtained for large delays. Note that the results in Figure 7-3 assume a user velocity of 3 km/h. For higher user mobility, the performance drops already at smaller backhaul latencies. However, the losses can be compensated by an improved link adaptation process as discussed by CT3.4 (see Section 7.3.1 and D3.3 [11]).

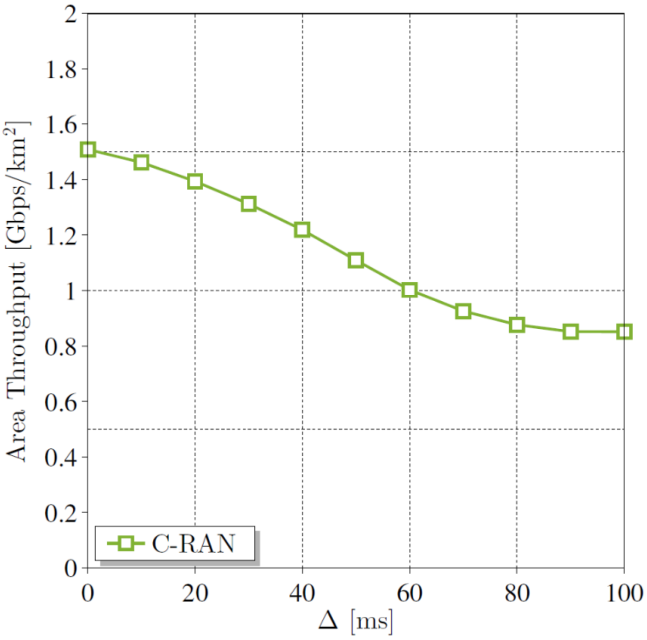

Figure 7-3: C-RAN Area Throughput as a function of the backhaul delay.

Backhaul Energy Consumption

A dense layout of small cells is necessary to satisfy the future traffic requirements and may result in an increase of energy consumption with higher operational costs for mobile network operators. Hence, all players in the wireless community have a tremendous interest for improving the energy efficiency at system level and are stimulating a large effort for finding innovative solutions. Current activities in this domain are mainly focussing on four research axes [20]: the definition of models to evaluate the energy efficiency of current access architectures, the design of new flexible hardware, the proposal of novel architectures based on dense small cells deployment, and the investigation of adaptive management schemes that adjust the network configuration with respect to service loads. However, most of previous studies have neglected the contribution of the small cell BH, which has a major impact on the overall power consumption.

Figure 7-4 compares the power consumption at a small cell and the associated backhaul (a microwave technology is considered here) [19] with respect to the small cell output RF power. These results show that nowadays BH power consumption is relevant and thus confirm that future energy saving schemes should consider jointly the impact of the small cell and BH on the overall network power consumption.

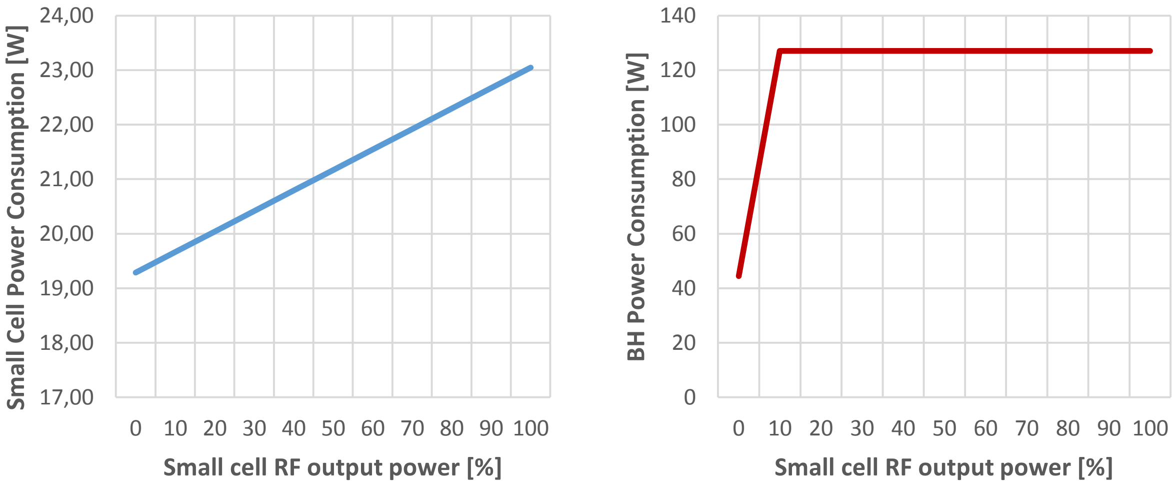

Figure 7-4: Small-cell and associated microwave BH power consumption [19].

In the joint BH/RAN design approach taken by iJOIN, the iNC monitors the network status and load, and runs the CT4.2 algorithms to optimise the overall energy consumption while ensuring that the network wide performance is not compromised. These algorithms are defined in D4.3 [14]. The Network-wide Energy Optimiser module (defined in D4.2 [13]) runs in the iNC and is in charge of taking network wide decisions about switching on/off physical nodes. It further ensures that UE traffic is properly routed by the nodes that are switched on at a given point in time. Thus, energy efficiency is achieved by minimising the number of network nodes switched on without affecting the UE QoS.

The SDN-controlled transport solution adopted in iJOIN takes into account the RAN status as well as the functional split level in operation on the different set of iSCs that compose the running veNBs. This allows for fine-tuning the forwarding within the backhaul network in order to adapt to the RAN requirements as well as to the backhaul status. Compared to solutions that are not capable to consider jointly the RAN and backhaul status, the iJOIN approach allows for full exploitation of the physical capacity of the network and dynamic adoption the existing conditions.

7.2 Joint RAN/BH Design and Operations

7.2.1 Joint RAN/BH Architecture

The transport plane of the veNB, i.e., the interconnection of the iSCs with the RANaaS instance, is realized by following a logically centralised approach based on SDN, where the iNC controls the network configuration (see Figure 7-5). This makes the backhaul network very flexible by tailoring and optimising it specifically for the requirements posed by the different functional splits.

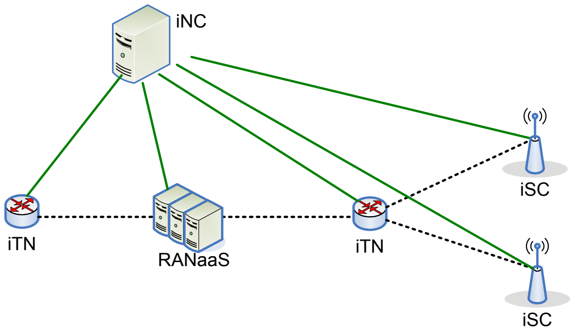

Figure 7-5: Transport network architecture

Figure 7-6 shows the message sequence chart implementing a joint RAN/BH functional design at the iNC and RANaaS instance. RAN measurements are first collected at the iSC level, then received and evaluated at the RANaaS instance (Step 1), which forwards relevant information to the iNC (Step 3). The information transferred to the iNC is typically related to the network load, QoS, and energy consumption. Additionally individual iTNs send BH measurements indicating the capacity and delay of the transport links to the iNC (Step 2). For instance, the latter can be used to concentrate the traffic in a limited number of nodes (see CT 4.2 [14]) and reduce the energy consumption (Step 4), or to change the number of hops between an iSC and the RANaaS to reduce the perceived delay (see CT3.1 [11]). RAN-aware BH orchestration is implemented at the iNC, which is in charge of updating the status of relevant iTN nodes (Step 5) according to the new configuration of the transport network and of informing the RANaaS instance about the new status of the backhaul (Step 6).

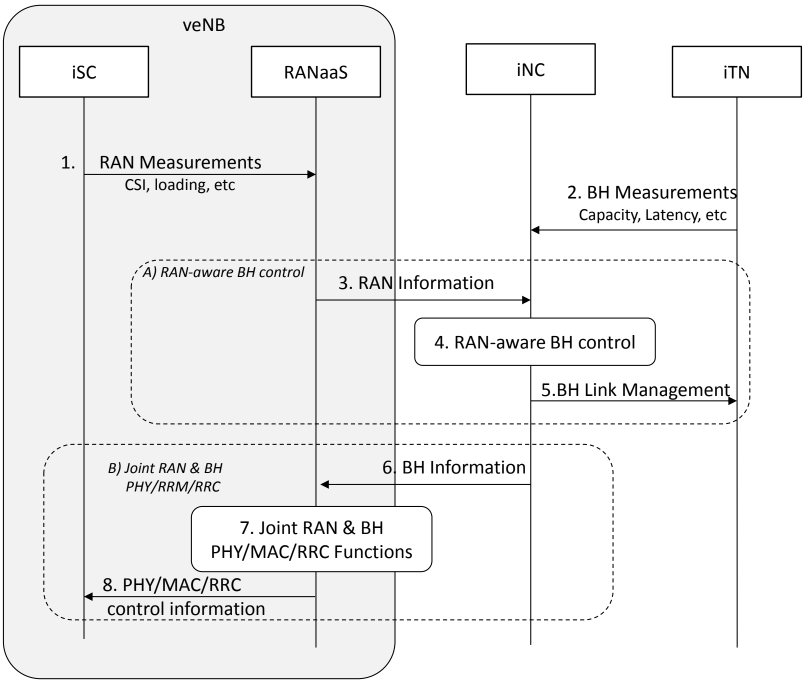

Figure 7-6: Message sequence chart for joint RAN & BH design

The RANaaS manager is responsible for performing management and orchestration tasks in support of veNB operations (7). These RANaaS operations can be optimised by jointly using RAN/BH measurements provided by iSCs (1) and the iNC (6), respectively. RAN measurements are typically related to the quality/strength of the radio links, the interference, and to the cell loads (i.e., the current radio resource usage and the composite available capacity). Moreover, the BH signalling may indicate the capacity and the delay of the transport links.

RAN measurements can be used to identify cells that suffer by excessive co-channel interference and they can be jointly used with the BH delay to setup robust ICIC mechanisms (see CT3.4 [11]) (Step 8). Moreover, measurements on the BH link quality may enable to increase the transmission robustness by introducing joint RAN/BH channel coding (see CT2.3 [8]). Additionally, information on the BH capacity can be used to identify bottleneck at some small cells, which need to share their load with small cells connected with high capacity BH links (CT3.2 [11]). In this case, the veNB implements load balancing procedures by advancing the handing over of the UEs between neighbouring cells or by changing the mobility parameters (Step 8).

7.2.2 Joint RAN/BH Network Energy Optimisation

Mobile operators typically dimension cellular networks in order to satisfy peak time traffic constraints and coverage requirements while limiting deployment costs. However, wireless networks are characterised by large spatio-temporal load variations, which are even more important in HetNets due to limited coverage of small cells. Specifically, previous studies have underlined that traffic daily profiles show a regular pattern during the day with low load periods during nights and early in the morning, medium loads during work-time, and high data rate in the late evening (see the left side of Figure 7-7) [21].

Since all kinds of eNBs are characterised by significant power consumption even at very low load scenarios, limiting the number of simultaneously activated eNBs can result in system-wide energy saving. Specifically, optimisation opportunities arise for adapting the network layout in the minute time scale according to the slow traffic variations (see the right side of Figure 7-7) due to (a) coverage overlaps stemming from heterogeneous or independent deployment of small cells, (b) large spatio-temporal load variations, and (c) load-dependent power consumption profiles of the wireless network nodes.

Beside this optimisation, faster adaptation can be implemented in the millisecond range by reconfiguring network parameters according to the fast traffic variations and fading. In particular, cell discontinuous transmission (DTX) has been proposed to dynamically deactivate most of the eNB hardware components, when data is absent in a given time frame [22].

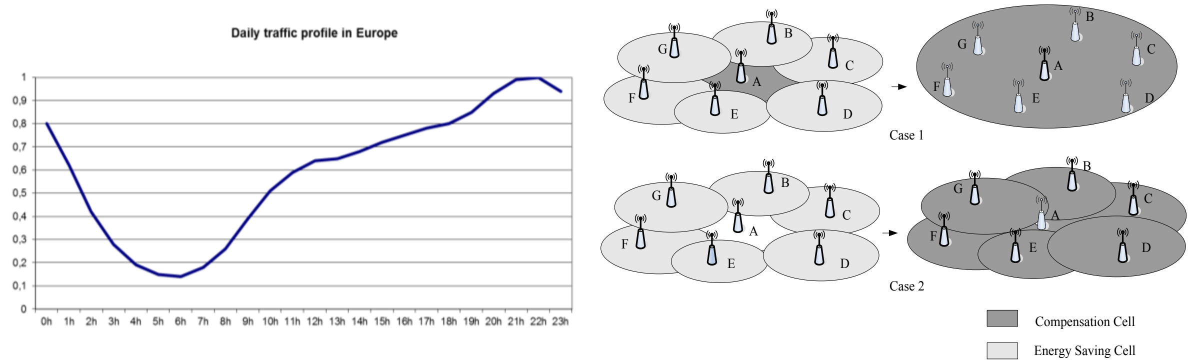

Figure 7-7: Normalized Daily Traffic profile in Europe [21] and Inter-iSC energy saving concept [23].

However, exploiting such opportunities must be done without violating agreed QoS constraints for users. First, energy saving mechanisms must avoid coverage holes, which can be experienced in the area where deactivated eNBs are located. Second, the more eNBs are turned off, the larger the traffic that active eNBs have to deal with, so the fewer the resources left for each served user. This may result in limited data rates for cell edge UEs and reduces the capability to support additional active UEs. Finally, by changing the network layout and concentring traffic on a limited number of eNBs, energy saving schemes may create spikes of interference, which affect the overall system performance.

To deal with these challenges, we propose a two stage mechanism that jointly uses inter-iSC energy saving and cell DTX. Specifically, our solution first identifies small cells that must be maintained active to guarantee coverage and average requirements associated to the different classes of data flows. Then, it routes the traffic of the dormant small cells towards the set of active small cells. Finally, it uses DTX to adjust the duty cycle of active small cells to momentary requirements of the data packets present in their queues. Details on these two steps can be found in [13] and [11].

Figure 7-8 shows the corresponding message sequence chart. As previously mentioned the inter-iSC energy saving mechanism is based on average measurements that permit the RANaaS to capture the access network load and coverage as well as the QoS requirements (1). In the same way, the iNC acquires BH link capacity measurements from the iTNs that are used to reconfigure the transport network path and re-route traffic across the network (2).

The RAN and BH measurements are jointly used at the RANaaS and iNC to dynamically reconfigure the transport and access networks as follows (3). First, the small cells to be deactivated are selected and associated UEs are handed-over to neighbouring active iSCs (4). Then, routing paths are updated and traffic is rerouted towards the serving cells (5). Finally, idle cells can be deactivated (6) without compromising coverage and rate constraints.

Although the latter mechanism increases the average load in the active small cells, there may still be sufficient degrees of freedom for further energy saving through cell DTX. In fact, as we mentioned, the momentary cell load may vary due to the inter-arrival time and the size of the packets, fast fading, and inter-cell interference. Furthermore, due to the service requirements, e.g. coverage, not all lightly loaded cell can be switched off.

In particular, cell DTX takes as inputs RAN measurements, which enables the RANaaS instance to estimate the cell throughput and the amount of data to be transferred in each time slot. The BH latency also affects the packet Time To Live; the higher the latency, the lower the time a packet can remain in the queue before it must be transmitted. This information is captured at the iNC through measurements done by iTNs which are then transferred to the RANaaS instance (7). Based on the received feedback and the queue status, in each sub-frame, the RANaaS instance implements the proposed DTX algorithm (8) to control the activity at small cells and associated BH links. When an activation is required, first a BH link activation request is sent to the iNC (9) that orchestrates the transport network (10) to enable the iSC activation (11) and successful data delivery to the corresponding small cells (12). Finally, each small cell locally implements resource allocation functions (13) such as scheduling and link adaptation according to the momentary channel quality information associated to the active UEs. Actual quantitative gains of iJOIN's energy-consumption improvements are discussed in Section 8.2.

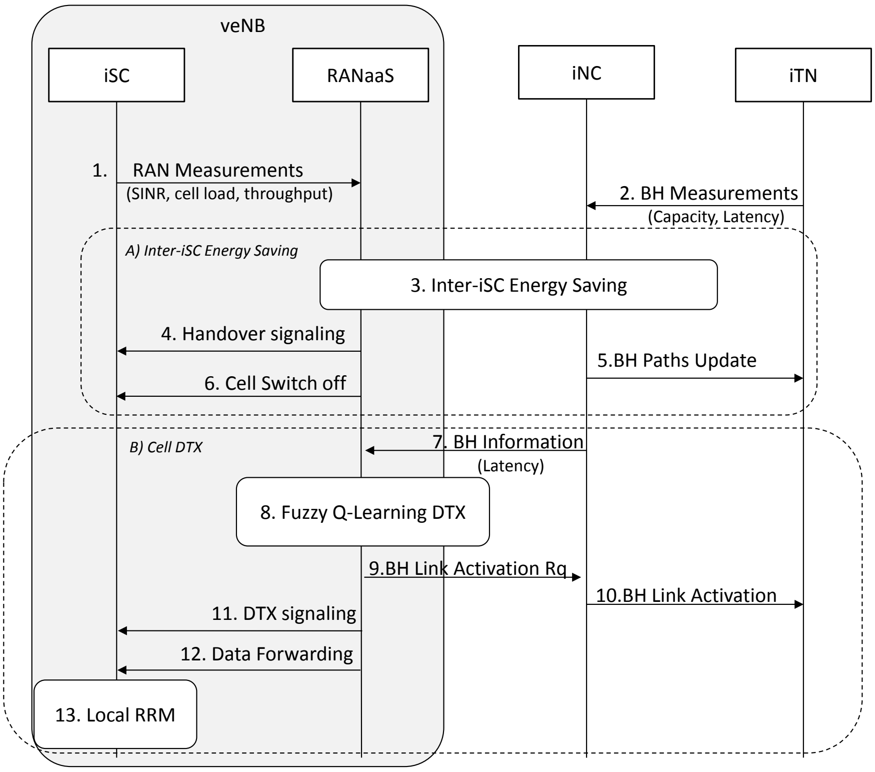

Figure 7-8: Message sequence chart for joint RAN & BH energy efficiency Optimization.

7.2.3 Joint RAN/BH Channel Coding

As discussed in previous sections, the BH can become a bottleneck for centralised processing, especially if BH technologies other than fibre have to be used. While, e.g., mmWave technology promises to be a low-cost alternative to fibre for the last mile, it has to be used as efficiently as possible to account for its lower data rate. Currently, BH and RAN links are treated as completely different network entities from a PHY-layer perspective. However, for highly centralised RAN protocol stacks, the BH only forwards the PHY-layer signal of the RAN channel. It is therefore beneficial to view them as part of the same system that can be designed jointly. Several joint coding techniques have been investigated to show the advantages and feasibility of this type of a joint RAN/BH.

Figure 7-9 shows the idea of joint error-protection coding. Current systems employ separate RAN and BH en-/decoders that operate using only channel quality information from their respective channels. However, due to the concatenated structure of the two channels, the RAN decoder can already account for the BH channel quality when setting the appropriate code rate. Additionally, channel quality information on the BH can also be forwarded to the RAN decoder, which improves its performance. Both techniques together can improve the area throughput by up to 110 % even when the BH link is limited due to its bad quality. Detailed investigations on this can be found in D2.3 [8]. These techniques have also been implemented using iJOIN's joint RAN/BH testbed where similar results were obtained [17].

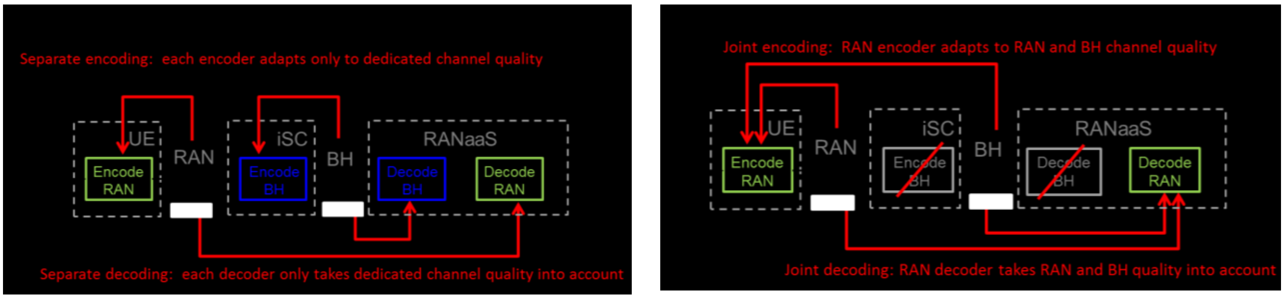

Figure 7-9: Separate coding for RAN and BH (left) compared to joint coding (right)

Similarly, a joint network-channel coding approach (JNCC) can be used to improve the overall area throughput if the BH throughput is limited. The concept is illustrated in Figure 7-10. The signals of multiple UEs or iSCs can be combined at aggregation nodes using network coding techniques before they are forwarded to the RANaaS. There, the individual signals can be extracted using low-density parity check (LDPC) decoders. As shown in D2.3 [8], this can improve the area throughput by up to 84 %.

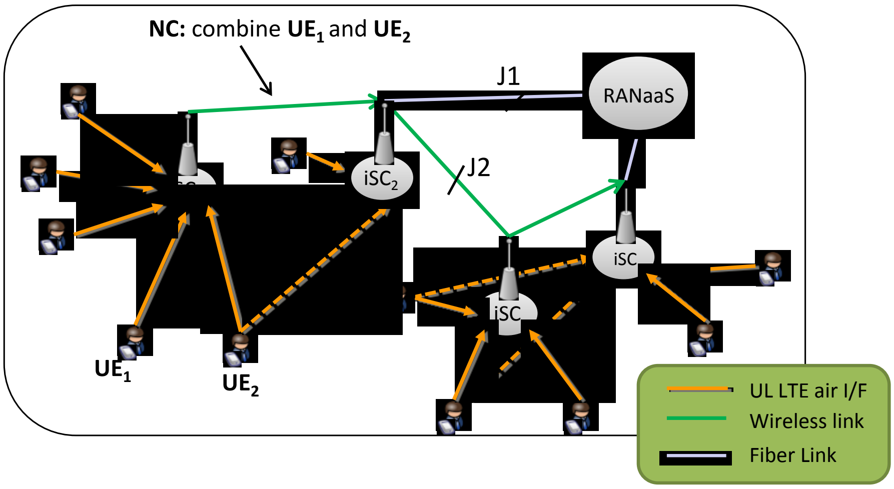

Figure 7-10: Scenarios for joint network-channel coding

7.3 Constraints and Requirements

7.3.1 BH Latency Constraints

One of the major challenges for computing link adaptation and/or scheduling centrally at the RANSaaS, is to deal with latency at the backhaul, especially when the coherence time of the channel is increasing due to high user mobility. In CT3.4 [11] we first introduced robust algorithms for both, link adaptation and scheduling. Those methods take into account that CSI available at the RANaaS is not precise, but affected by several sources of impairments, such as, noisy pilot reception for estimating the channel, quantization effects and delays. The last aspect is of special interest for architectures with centrally processing when the backhaul is affected by delays. The robust algorithms of CT3.4 incorporate the statistics of the CSI impairment into the processing.

In addition, CT3.4 presents a multi-stage scheduling approach, which splits the scheduling into a central and a local part as shown in Figure 7-11.

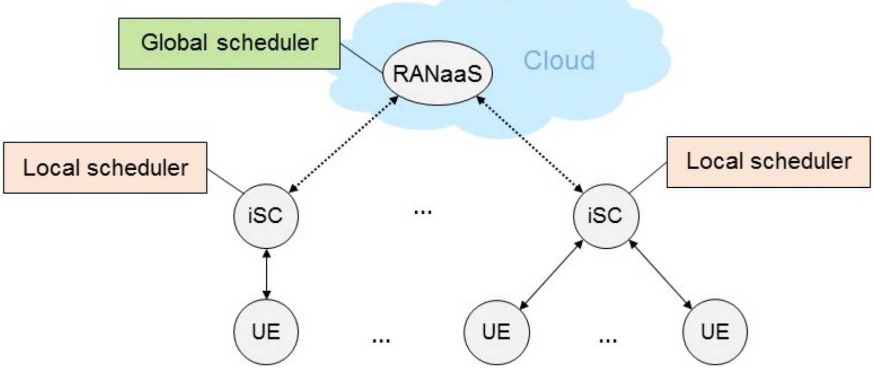

Figure 7-11: Semi-deterministic, hierarchical multi-stage scheduling [11].

The former one is processed at the RANaaS, while the latter one is places at the iSCs. Such split allows to exploit global channel knowledge for performing coordinated scheduling of multiple iSCs. However, due to backhaul delays, the scheduling decisions are based on outdated CSI. The multi-stage scheduling approach of CT3.4 allows to update the central assignment at the iSCs based on local CSI, which is less outdated. In this regard, the global scheduler decides which iSC is transmitting and which is not. With the local update, the user selected to be served at a particular time slot can be changed at the iSC. Consequently, with this approach coordinated scheduling can still be applied in a scenario where the backhaul latency is in the range of the channel coherence time.

As shown in Figure 7-12, with the algorithms presented in CT3.4 up to 57% gains can be achieved compared to an LTE setup, which implements a reuse 1 system with round robin scheduling.

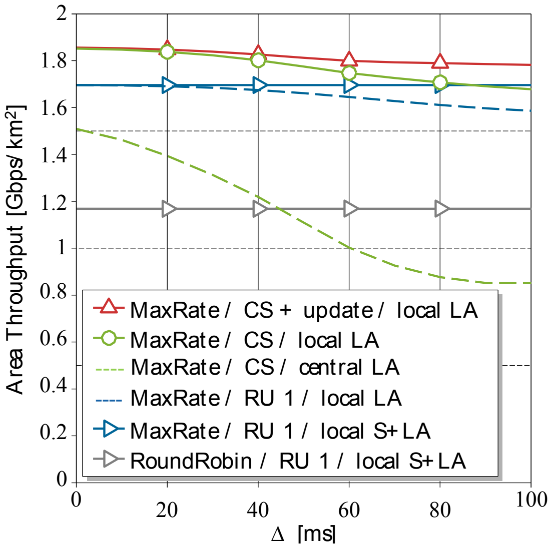

Figure 7-12: Area throughput as a function of the backhaul delay for a user velocity of 3 km/h [11].

7.3.2 Congestion Control

When considering the joint backhaul and RAN design, it should be noticed that congestion in the backhaul will have different consequences depending on the actual functional split affected by the congestion. Congestion usually results in some form of loss of information blocks (frames, packets), and while the loss of one or more of them may produce just a degradation of the quality of service for some functional splits, for others it may have catastrophic consequences, affecting multiple users and even causing the loss of the communications. In this sense, taking into account the potential effect of congestion in the backhaul depending on the actual functional splits is a necessary constraint that should be taken into account. On top of this, it is required a congestion control mechanism that can support over the same infrastructure flows from different functional splits and handle their different requirements in an optimized way. If this mechanism is not available, then the only solution is either overprovisioning so congestion cannot happen or having separate transport infrastructure for different functional splits.

The objective of CT 4.4 [14] is to provide an adequate technical solution to address the above described challenge of congestion control in the transport nodes that support RANaaS and flexible functional split. In fact, actual congestion control mechanisms are not able to provide the required support for the specific iJOIN requirements, i.e., either variants of the TCP based congestion control or new ones developed for the data centre realm.

For a congestion control mechanism to be efficient, it should be able to handle in an effective way different kinds of flows associated with different functional splits, which may present different requirements in terms of capacity or maximum latency. An example of this situation is shown in Figure 7-13.

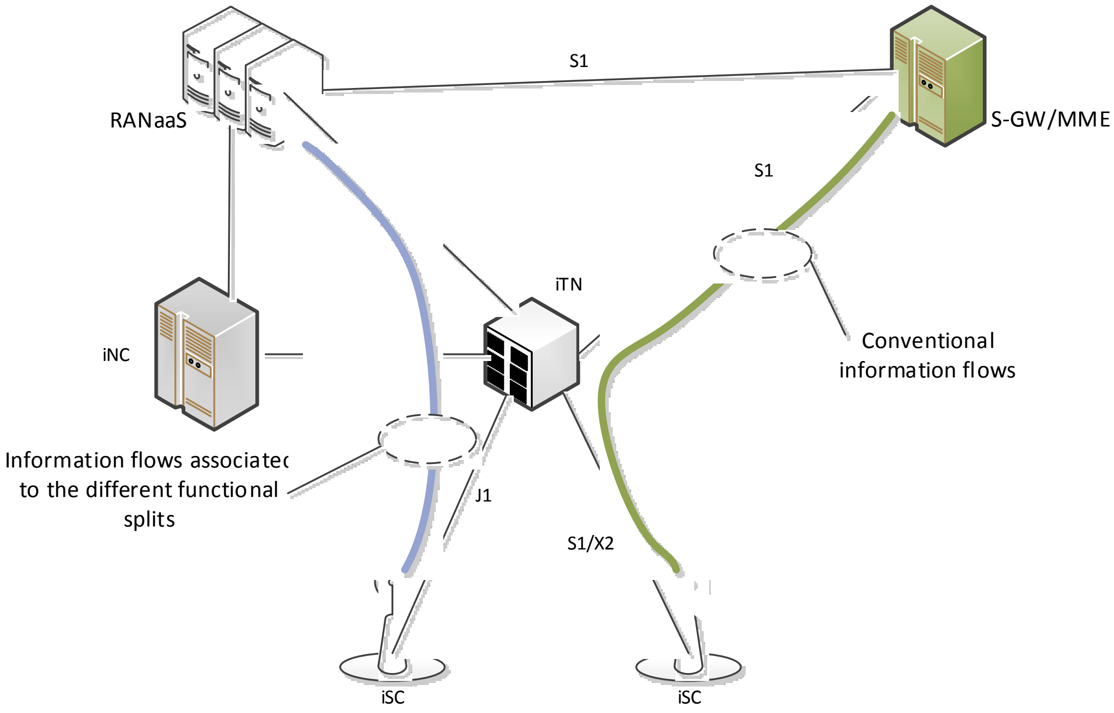

Figure 7-13: Congestion Control Example

It seems clear that coping with this kind of situations requires new technical solutions, closer to those employed in data centres than traditional congestion control solutions. The flows that can traverse an iTN have been classified, for the purposes of controlling congestion, into two main sets:

- Variable bit rate flows (VBF), where information is sent based on the generation patterns of the users.
- Constant bit rate flows (CBF), associated with functional splits that generate information blocks with a fixed frequency.

The characterisation of these flows, especially the variable bit rate ones, is a complex task as the flows are the result of the aggregation of the traffic for or from different users. This traffic may also have been originated from several sources, e.g., different applications running in the user device, some of them possibly in the background. It can be assumed that CBFs will require, on top of a higher transport capacity, very low latency and packet drop probability when compared with VBFs.

Congestion control effectiveness can be evaluated looking at a number of metrics that characterise the effectiveness of the proposed technology (see D4.3 [14] for more details).

The solution proposed by iJOIN is based on the virtual queue concept, which focuses on eliminating queues at all switch ports of the transport node, so latency is reduced to the minimum. The concept is motivated by the idea that it is possible to eliminate buffering and queueing delay by detecting congestion based on the link utilisation approaching its capacity, rather than the queue occupancy. In this sense, the virtual queue represents an imaginary flow whose unused capacity can be used to accommodate traffic increases without queue build-up. The implementation of this CT is based on several functional entities that can be located at the iTN or distributed between the iTN and the iNC (see D4.2 [13] for more details).

Figure 7-14 shows results from different sets of variable bit rate flows traffic intensity of 1.5 Gbit/s

- Points marked as "+" correspond to a policy that drops packets but does not activate an Explicit Congestion Notification (ECN).
- Points marked as "o" implement a policy with a double threshold in the virtual queue, a lower one that activates ECN and a second one that activates packet dropping.
- Points marked as "x" corresponds to a congestion control based only in ECN

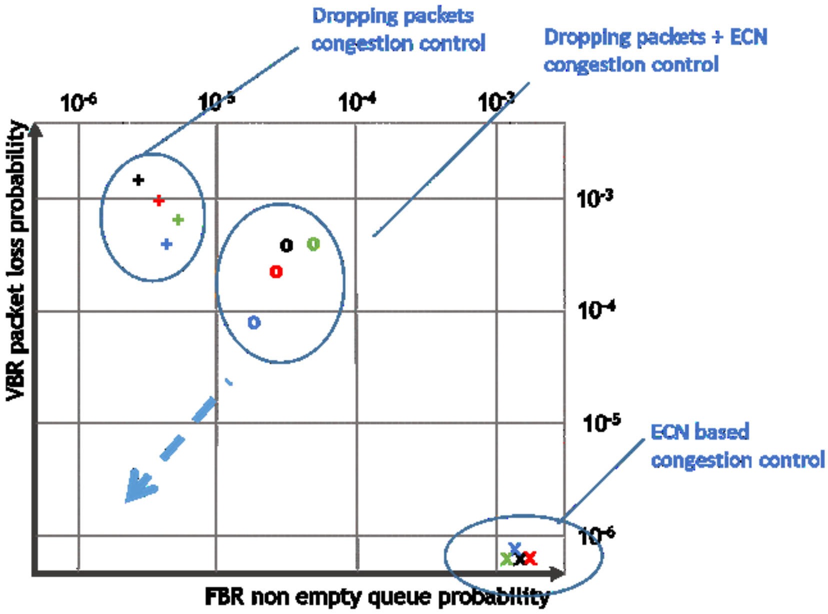

Figure 7-14: Uplink non empty egress port queue probability vs VBFs' packet loss probability

We can see that ECN based congestion control does not benefit from any change in the CT configuration parameters in order to reduce the probability of non-empty queue for the fixed rate flows. This is due to the fact that the mechanism is not fast enough. It is possible to improve the variable bit rate dropping rate acting on the CT configuration parameters without sacrificing the probability of finding the queue empty by the fixed rate flows. Specifically, the use of additional information such as the occupation of other queues in the network (that would be available for a centralized implementation of the CT) may help to avoid dropping an excessive number of packets in successive iTNs. Also, it has been verified that a careful selection of the dropping probability may also be helpful.

8 iJOIN System Evaluation

8.1 Evaluation methodology

The two iJOIN key concepts are evaluated with respect to its four key objectives area throughput, energy efficiency, utilisation efficiency, and cost efficiency which were defined in the previous deliverable [15]. In order to demonstrate the effectiveness, we evaluate the two main operational objectives area throughput and energy efficiency based on four scenarios:

1. **Stadium scenario, Split A**: In this case, we apply the scenario described in Section 5.5.1 as well as the preferred functional split A (see Section 6.2) which requires very high-throughput and low-latency backhaul as in the case of CPRI.
2. **Wide-area scenario, Split A**: In this case, we apply the scenario described in Section 5.5.2 as well as the preferred functional split A. By contrast to the stadium scenario, this scenario covers a large area and is subject to mobility. Furthermore, Split A represents a dark-fibre solution and we will demonstrate how iJOIN technologies can improve area throughput and energy efficiency over existing technologies.
3. **Wide-area scenario Split B**: Again, the wide-area scenario is considered but also backhaul with higher latencies, resulting in preferred functional split B. We will use this scenario to demonstrate how iJOIN can maintain a major part of the area throughput and energy gains compared to the previous scenario. In addition, this scenario will show how iJOIN allows for exploiting these gains in scenarios in which centralisation was not feasible with state-of-the-art technologies and therefore these gains could not be exploited.
4. **Wide-area scenario Split C**: Similar to the previous scenario, this scenario shows how to exploit centralisation gains in scenarios which so far did not support any PHY centralised processing. Compared to the previous scenario, the backhaul requirements are further reduced as mainly legacy backhaul is considered as well as scenarios in which backhaul deployment is very difficult (e.g. NLOS backhauling).

These four scenarios will allow for showing the effectiveness of the iJOIN system concept. In addition, each work package performed a detailed evaluation of all four common scenarios and different functional splits which are described in [8] [11] [14]. In order to allow for the comparison of results and to facilitate the integration of results across multiple work packages, the system assumptions as listed in D5.2, Table 7-1 [15] for the stadium scenario and D5.2, Table 7-2 [15] for the three wide-area scenarios are applied.

Based on the individual evaluations in [8] [11] [14], each work package identified candidate technologies which can be applied to the four scenarios. Each of the selected candidate technologies is then analysed in the context of the previously explained scenarios and using the same parameter sets. The performance improvements of the selected candidate technologies are then combined in order to determine the achievable benefits of the iJOIN system concept.

In addition, Section 8.2.5 presents a system-wide evaluation of iJOIN’s cost-efficiency, and Section 8.2.6 presents a system-wide evaluation of iJOIN’s utilisation efficiency.

8.2 System-wide Results

8.2.1 Scenario 1: Stadium and Split A

Table 8-1: CTs evaluated for area throughput in Scenario 1

Objective	CT	Short description	Gain
Area Throughput	CT 2.2 + CT 3.7	Multi-point Turbo-Detection (UL)	56% (1UE/iSC) 60% (5UEs/iSC)
	CT 2.4	Centralised CoMP (DL)	30%
	CT 2.6	Fronthaul Compression (DL and UL)	100%

The first scenario is the stadium scenario (Section 6.2) and preferred functional split A where all digital processing at least after the computation of the FFT is centralised. Hence, this implies similar FH requirements as CPRI with the comparison that CT 2.6 is able to significantly reduce the required fronthaul data rates (factor of 0.5). Therefore, under a given fronthaul data rate constraint, CT 2.6 is able to either support twice as many radio access points for the same per-cell throughput or double the supported radio access throughput for a constant number of radio access points.

In the downlink, CT 2.4 (fully centralised CoMP) is the means of choice to improve data rates by up to 30%. In D2.2 [4] and D2.3 [8], the optimal cluster-size for one virtual eNodeB under a given fronthaul constraint has been analysed in the context of CT 2.4. This is of particular relevance for the stadium case where depending on the actual user traffic profiles, different requirements on RAN and fronthaul are imposed and therefore the optimal veNB size may change over time.

In the uplink, CT 2.2 (multi-point turbo-detection) and its matching MAC protocol definition in CT 3.7 are applicable. CT 2.2 exploits inter-cell interference in order to improve the SINR. In the case of very high data rate demands which result in full spectrum usage by each radio access point, CT 2.2 is the only way to sustain the require data rates as inter-cell interference cannot be avoided.

In order to obtain the combined gains, we assume that in the best case gains of compatible CTs are independent of each other. CT 2.2/3.7 improves the performance by 56% (factor 1.56) and CT2.6 by 100% (factor 2). Hence, combining both CTs may result in a performance gain of up to 212% (factor 3.12). Similarly, in the downlink a performance gain of up to 160% may be achieved.

Note that this gain is applied on top of the expected network densification by a factor of 10 or higher. The described candidate technologies are enablers to sustain the increase of system throughput with the increased network densification, which may not be possible with existing technologies. Therefore, the expected overall area throughput improvement highly depends on the system and data rate demand assumptions.

8.2.2 Scenario 2: Wide-Area and Split A

Table 8-2: CTs evaluated for area throughput and energy efficiency in Scenario 2

Objective	CT	Short description	Gain
Area Throughput	CT 2.2 + CT 3.7	Multi-point Turbo-Detection (UL)	51% (2UEs/iSC) 65% (10UEs/iSC)
	CT 2.4	Centralised CoMP (DL)	30%
	CT 2.6	Fronthaul Compression (DL and UL)	100%
	CT 3.5	RRM for interference coordination (DL)	18%
Energy-Efficiency	CT 3.3	Cell Discontinuous Transmission	70%
	CT 4.2	NW wide energy optimization	35%

The second evaluation scenario uses the Wide-Area Coverage scenario described in Section 5.5.2 and again the preferred functional split A. As before, we again apply in downlink CT 2.4 (centralised CoMP), in uplink CT 2.2 (multi-point turbo-detection), and CT 2.6 for backhaul compression. By contrast to the stadium scenario, we additionally evaluated CT 3.5 (inter-cell interference coordination). CT 2.4 and CT 2.6 are again the means of choice if the available spectrum is fully used and therefore strong inter-cell interference is caused.

By contrast, CT 3.5 does not exploit interference but rather avoids strong inter-cell interference. As spectrum utilisation per radio access point decreases, CT 3.5 can exploit additional degrees of freedom for its resource management in order to avoid interference. However, CT 3.5 is less computationally intense and less susceptible to imperfect channel state information. Hence, CT 2.4 and CT 3.5 would not be applied at the same time but are used depending on the actual data rate demands.

The overall improvement of area throughput in uplink is up to 230% and in downlink up to 160% based on the results above. Note that also these gains are applied on top of the expected network densification by a factor of 10 or higher.

Energy-efficiency has been the main objective for CT 3.3 and CT 4.2 which were evaluated for this scenario. CT 3.3 implements discontinuous transmission on a short timescale while CT 4.2 optimises network-wide energy-consumption by turning off base-stations. Both CTs have an inherent interaction because CT 4.2 does reduce the number of active base-stations and potentially re-routes traffic to otherwise lower utilised base-stations. CT 3.3, by contrast, exploits the fact that a base-station cannot be fully turned off because some user terminals cannot be re-assigned to other base-stations. Hence, CT 4.2 reduces the degrees of freedom for CT 3.3. In the best case, gains of both CTs would multiply, i.e. CT 4.2 turns off as many base-stations as possible without incurring significant additional traffic to enabled base-stations. This is the case if coverage areas do not (significantly) overlap or if only very few user terminals are attached to the network. For the remaining under-utilised base-stations, CT 3.3 would then further reduce the energy-consumption on a short timescale. Hence, the overall energy-efficiency gain is up to 79%.

Both, CT 4.2 and CT 3.3 can be combined with CT 2.2, CT 2.4, and CT 2.6. However, as more and more base-stations are turned off, also the inter-cell interference pattern changes which impacts CT 2.2 and CT 2.4 (Section 8.2.7 provides more details on this trade-off).

8.2.3 Scenario 3: Wide-Area and Split B

Table 8-3: CTs evaluated for area throughput in Scenario 3

Objective	CT	Short description	Gain
Area Throughput	CT 2.1 + CT 3.8	In-Network Processing (UL)	130%
	CT 2.3	Joint Channel-Network Coding	30%
	CT 2.5	Hierarchical CoMP (DL)	30%
	CT 2.7	Joint RAN/BH Coding (UL)	110%
	CT 3.5	RRM for interference coordination (DL)	18%

The next scenario is again a wide-area scenario but with the preferred functional split B where only user-specific functionality is centralised, i.e. forward error correction and anything above. In the downlink, we apply CT 2.5 (Hierarchical CoMP) which splits the pre-coding process into two stages of which one is centralised and one is performed locally. For the given scenario, CT 2.5 achieves gains of up to 30%. Alternatively, again CT 3.5 (inter-cell interference coordination) is applicable and achieves gains of up to 18%.

In the uplink, our first option is to apply CT 2.1 (In-Network Processing) and its matching MAC protocol defined by CT 3.8 together with CT 2.7 (Joint RAN/BH Coding). CT 2.1 and CT 3.8 improve area throughput by up to 130% and CT 2.7 by up to 110%. The main performance benefits of CT 2.1 originate from an improved signal detection performance. By contrast, the performance benefits of CT 2.7 stem from an improved joint encoding over RAN and BH. Hence, the performance benefits of both can be combined to an expected uplink throughput improvement of up to 380%.

The energy-efficiency evaluation for this scenario follows along the same lines as for the previous scenario with preferred functional split A. Again, the application of CT 3.3 and CT 4.2 can be combined with the CTs considered in this scenario but they potentially affect the performance gains due to a change of the inter-cell interference pattern.

8.2.4 Scenario 4: Wide-Area and Split C

The last scenario that has been investigated is again based on the wide-area scenario but applies preferred functional split C where MAC layer functionality and anything above is centralised. Due to the lower degree of centralisation, also the degrees of freedom for inter-cell cooperation decrease. Therefore, on physical layer we only consider CT 2.2 (single-point turbo-detection) and its associated MAC protocol in CT 3.7. In downlink, we consider again CT 2.5 (hierarchical CoMP) which allows for splitting up the pre-coding process. The area throughput gains achievable with these CTs are up to 26% in uplink and up to 25% in downlink.

The main reason for the lower gains of physical layer technologies in this scenario is the high backhaul latency and low backhaul throughput which both reduce the degrees of freedom for joint processing in uplink and downlink. By contrast, WP3 investigated technologies on MAC layer which rather avoid and coordinate inter-cell interference. In downlink, CT 2.5 exploits inter-cell interference through joint transmission while the investigated MAC layer approaches rather avoid interference. Hence, it should not be combined with any other scheduling approach. CT 2.2 is the only uplink CT evaluated for this scenario.

Table 8-4: CTs evaluated for area throughput in Scenario 4

Objective	CT	Short description	Gain
Area Throughput	CT 2.2 + CT 3.7	Single-Point Turbo-Detection (UL)	20% (2UEs/iSC) 26% (10UEs/iSC)
	CT 2.5	Hierarchical CoMP (DL)	25%
	CT 3.1	Joint RAN/BH link-scheduling	140%
	CT 3.4	Semi-deterministic Scheduling	57%
	CT 3.9	Cooperative scheduling and interference control	20%

CT 3.1 (joint RAN/BH link scheduling) jointly schedules resources on backhaul and radio access network in order to avoid bottlenecks. This is particularly important in the case of backhaul with strong throughput constraints, e.g. multi-point mmWave backhaul technology. CT 3.4 (semi-deterministic scheduling) provides most of its gains in scenarios with mobility where only imperfect channel state information is available, i.e. the proposed CT is very robust. Finally, CT 3.9 (cooperative scheduling and interference control) is very flexible, uses only long-term channel state information, and it is applicable to strongly limited backhaul. Given the scenario, area throughput gains in downlink of 20%-140% are achievable, and up to 26% in uplink.

8.2.5 Cost-Efficiency Study

Table 8-5: Parameters applied to the cost-efficiency study

Type of cost	DRAN	Cloud-RAN
Macro base station	$50k	$25k
Micro base station	$20k	$10k
Microwave BH	$50k per link plus $5k per kilometre	
Optical fiber BH	$5k per link plus $100k per kilometre	
Data center		$40k
Server blades		$20k each (see [37])

The cost-efficiency evaluation of the iJOIN system has been performed based on the stochastic geometry framework described in D5.2 [15] and D4.2 [13]. This framework considers four different network domains: the radio access network with a very dense small-cell deployment, an overlaid macro-cellular deployment, the backhaul network which may be composed of optical fibre and wireless backhauling technologies, and the data centre layer which performs centralised functionality. The framework can be evaluated for both iJOIN's RANaaS concept as well as distributed RAN (DRAN) by substituting the appropriate component cost values. Distributed RAN refers to a conventional implementation where all RAN processing is performed by base stations. Table 8-5 shows an exemplary budget for an iJOIN deployment that uses the preferred functional split B. It compares the costs for DRAN, fully centralised RAN, and RANaaS where centralised processing is performed on commodity hardware. In our example, we assume 170 active users per km^2, an average traffic demand per user of 10Mbps, and a mix of 50% microwave and 50% optical fibre technology to connect base stations with backhaul nodes.

Figure 8-1 shows the resulting CAPEX for DRAN, Full CRAN, as well as for three different configurations of RANaaS over different data centre densities. It is important to note that one data centre may consist of only a few server racks at an existing point of presence within the mobile network. This reduces the OPEX because no additional site rental is necessary. Furthermore, small data centres promote greater failure resilience and they reduce the traffic within the metropolitan transport network. Therefore, considering Figure 8-1, a density of one or two data centres per square kilometre appears realistic in a very dense urban small-cell deployment. If we further increase the density of small-cells, e.g. due to higher data rate demands and user density, then the cost effectiveness of RANaaS would increase even further as the exploited multiplexing gains also increase (similar to the over-provisioning of distributed RAN) and the cost-reduction per base station becomes more dominant.

The results show that RANaaS based on the applied RAN functional split can be more cost efficient than a DRAN implementation. The actual benefit may depend on the scenario, parameterisation, and actual traffic demand. Additionally, the architecture presented allows for taking advantage of reduced OPEX due to lower maintenance costs on site as well as easier management of RANaaS using standard IT management mechanisms.

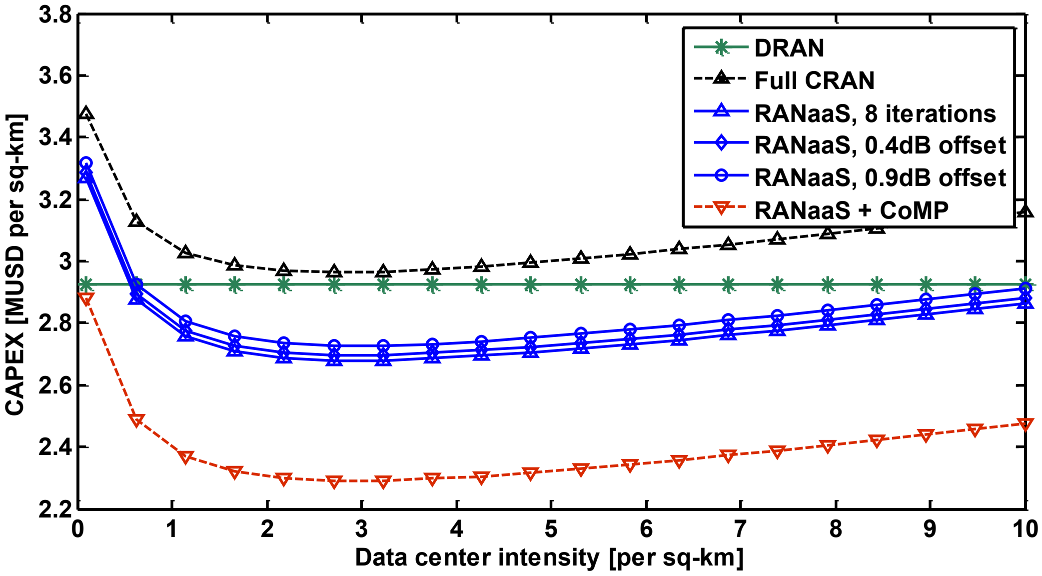

Figure 8-1: Cost-efficiency results for iJOIN concept

As seen from the three RANaaS curves in Figure 8-1, the change of link-adaptation parameters is fairly cost neutral (see for more details D3.3, Section 4.6 [11]). Using higher link-adaptation thresholds and therefore, lower decoder complexity, reduces the achievable throughput. Hence, additional base stations must be deployed in order to satisfy the user demands. On the other hand, fewer data processing resources are necessary which reduces the costs for data centres and these effects counteract each other. However, there is one significant performance advantage: if low-complexity decoders are applied (e.g., a turbo-decoder with fewer iterations), the processing time per codeword is reduced and so is the software-latency (see also the discussion in D2.2, Section 3.2 [4]). Software-latency refers to the latency induced by the RAN processing software running at the RAN instance. This latency is very critical as it easily becomes a bottleneck for RAN protocol processing.

Finally, Figure 8-1 shows the performance for RANaaS and CoMP. In this case, we assume that the spectral efficiency has been improved by 30% as shown by CT 2.4 and that the network density can be reduced accordingly. Although a practical deployment may not be able to exploit all these degrees of freedom, it still shows the impact of CoMP on the cost-efficiency of the system. Given these assumptions, the CAPEX of the system could be reduced by up to about 25% compared to a DRAN implementation.

8.2.6 Utilisation-Efficiency Study

Utilisation can be evaluated in three different domains: the radio access network, the backhaul network, and the data centre. It measures the expected usage of available resources such as available capacity or processing hardware. A detailed introduction of utilisation efficiency has been given in D5.2, Section 7.2.3 [15]. Based on this definition, utilisation efficiency has been exemplarily evaluated in all three domains.

First, we consider the utilisation efficiency in RAN and backhaul network. CT 3.2 introduces a novel joint RAN and backhaul connection control. Using this technology, it is possible to re-assign user terminals to different cells in order to turn off base stations as well as backhaul nodes. Hence, this technology improves both energy efficiency and utilisation efficiency.

Table 8-6: CTs evaluated for utilisation efficiency

Objective	CT	Short description	Network Scope	Relative Gain
Utilisation Efficiency	CT 3.2	Joint RAN/BH load balancing	Radio access and backhaul network	50%
	CT 3.6	Joint RAN/Cloud scheduling	RANaaS instance	200%
	CT 4.1	Distributed Mobility Management	Transport network (RANaaS → gateways)	45% (iLGW in BH) 60% (iLGW in RAN) 75% (iLGW in BH&RAN)

A full description of the technology is given in D3.2, Section 4.2 [10]. CT 3.2 has been evaluated using the iJOIN baseline assumptions and the results of this evaluation are presented in D3.3, Section 4.2 [11]. Figure 8-2 shows the results over different user densities for CT 3.2, and two small-cell deployments with different cell range extension bias (0dB and 3dB). Furthermore, the backhaul is assumed to be limited by 120 Mbps in order to have a basis for the computation of the utilisation. We can see that the combined utilisation of RAN and backhaul improves from about 30-60% to 50-90% if CT 3.2 is applied.

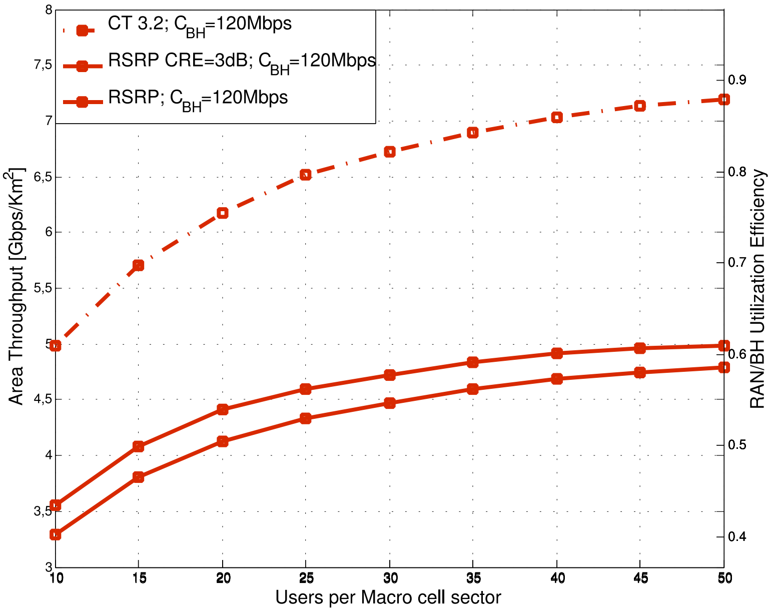

Figure 8-2: Evaluation of RAN and BH utilization using CT 3.2 [11].

Next, we consider the utilisation of data centres for a given number of centralised base stations. Based on the framework presented in D3.2, Section 4.6 [10] and [35], we are able to predict the amount of computational resources required to operate a RAN protocol stack. In this investigation, we assume preferred functional split B where forward error correction decoding is centralised and no statistical multiplexing gain is exploited, which would improve the results even further. Using the framework in [35] and the measured relative overhead for higher layer processing discussed in [36], we can determine the utilisation efficiency at the RANaaS instance.

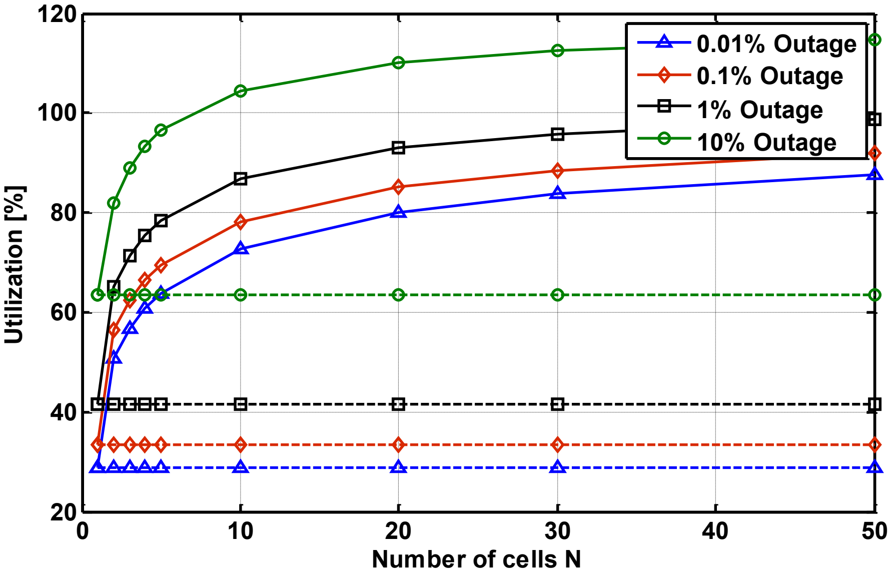

Figure 8-3: Utilization efficiency of RANaaS instance

In order to dimension the RANaaS, we assume that sufficient resources are provided in order to process signals from any arbitrary base station with probability $1-\epsilon$, where $\epsilon = 10\%, 1\%, 0.1\%\ and\ 0.01\%$. CT 3.6 presents a joint scheduling technology for RAN and RANaaS instance which avoids any computational outage at the cost of a marginal system performance degradation. For instance, if computational outage probability of 10% is used to dimension the system, then in 10% of the cases CT 3.6 must be active to avoid a computational outage. In order to achieve this, the link-adaptation of some base-stations is slightly modified which causes an average performance degradation of 0.3%. Figure 8-3 shows numerical results for utilisation efficiency of the RANaaS instance and, as a comparison in dashed lines, for a local RAN implementation. We can see that depending on the target computational outage, a local implementation runs at 30%-60% utilisation of its processing resources (although a practical system would rather be optimised for very low computational outage resulting in utilisation below 30%). By contrast, a centralised implementation increases the utilisation to at least 90%. In the case of $\epsilon = 10\%$ we can further see that utilisation achieves values above 100%, i.e. the system would run permanently at full load. As mentioned, in this case CT 3.6 would slightly reduce the system performance by 0.3% in order to avoid any computational outage but maintain full utilisation.

Finally, CT 4.1 evaluated the utilisation of the transport network between data-centres and gateways, i.e. links between RANaaS instance and radio access points are not considered by CT 4.1. Therefore, this CT can be combined and operated jointly with CT 3.2 and CT 3.6. CT 4.1 has been evaluated in three different configurations of the wide-area scenario and the results are reported in D4.3, Section 7.1 [14]. The three scenarios differ in the considered backhaul connectivity, i.e. the first scenario considers optical fibre, the second scenario considers additional mmWave links, and the third scenario considers sub-6GHz backhaul technologies and mmWave links between iTNs.

8.2.7 Combined Results

Figure 8-4 shows the relationship of the described results (average of uplink and downlink results) to the system design concept introduced in Section 4. Note that the figure does not show the gains that are achievable by network densification itself, which would come on top. Furthermore, it shows the **average of uplink and downlink improvement** in the case of area throughput. Each marker in the figure represents one of the scenarios described in the previous sections.

The system design can be divided into two parts: a planning part and an operational part. In the former, we first define a minimum area throughput that must be achieved by the system. Depending on the scenario and assumptions, different minimum network densities result. Given the CTs discussed above, improvements of up to 300% can be achieved which also reduces the necessary network density. Based on the available technologies in RAN and BH, a cost-efficiency optimisation is performed. As shown before, iJOIN's concepts are able to reduce the cost by at least 10% (if no cooperative technologies are deployed) and up to 25% if CoMP is deployed. Then, given a minimum area throughput and deployment strategy, also the utilisation can be maximised in the system, e.g. by dimensioning data centres accordingly.

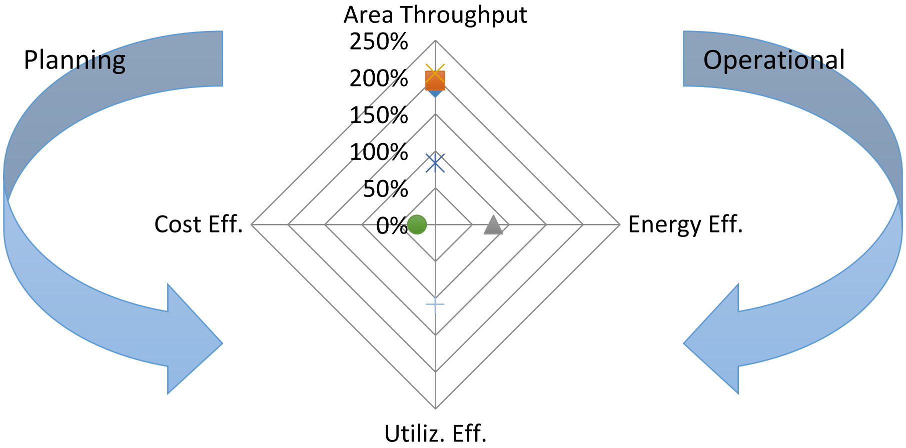

Figure 8-4: Relation of results to system design concept

In the case of the operational design, we again define a minimum area throughput under the given number of user terminals and service requirements. For the given data rate demand, the energy-consumption is reduced without violating the data rate constraint. Apparently, the number of base-stations that can be turned off to save energy also depends on the spectral efficiency per base-station. Hence, technologies that enable higher spectral efficiency also allow for higher energy-savings. Given the results above, these energy-savings may be up to 80%. Finally, also the utilisation of base-stations is maximised by means of congestion avoidance in all three network domains, i.e. RAN, backhaul, and data centre. Our results show that utilisation may be improved by 45%-108%.

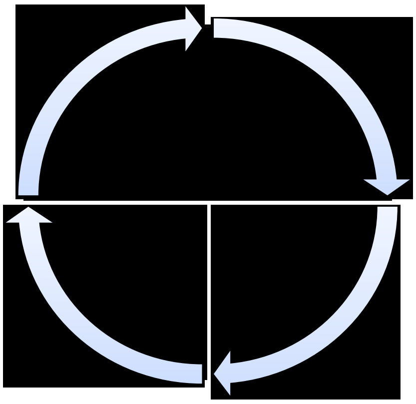

Figure 8-5: Actual achievements regarding iJOIN key objectives

Using these results, we can derive the actual achievements regarding iJOIN's key objectives (see Figure 8-5) and compare them with the initial target shown in Figure 4-3. The initial goal for area throughput has been an improvement by a factor of 50-100. The results for area throughput obtained through iJOIN technologies are summarised in Table 8-7. By averaging uplink and downlink results, we can conclude that iJOIN's CTs achieve gains by a factor of up to 62 under the assumption that the network density increases by a factor of 10 and four antennas at each small-cell (compared to two in the baseline assumption). This improvement is slightly lower than the target improvement. In any case, we believe that this is a very remarkable result, and shows that very high gains are obtained from iJOIN's key concepts. It is worthwhile noting that the previous does not take into account additional effects from propagation characteristics, e.g. in the case of a very dense network the line-of-sight probability increases which may further improve the performance. However, there are no reliable channel models available yet for highly densified networks, and therefore this investigation was out of scope for iJOIN.

Before, we have shown that the energy-efficiency can be reduced to about 20% of the baseline architecture. However, these 20% have been achieved before further network densification by a factor of 10. Note that for a densification by a factor of 10 and path-loss exponent 3, the overall consumed transmit energy is reduced by a factor of 30 (if the minimum SNR is used as normalisation). In this regime, the fixed power consumption will be significantly higher than transmit energy, i.e., the latter would be negligible. According to the model in D5.2 [15], in the case of iSCs, the transmit energy is in the same order as the fixed power consumption. Hence, the energy consumption would be further reduced by a factor 2, which results in the 10% shown in Figure 8-5 being only slightly higher than the targeted 5%. Furthermore, we expect that additional energy gains could result from more energy-efficient hardware which will be available in the future, and therefore it is reasonable to expect that in that case the target 5% will be very likely achieved. However, there is no reliable data yet available on hardware platforms that are used in 2020 and therefore a quantitative evaluation of this additional gain was out of scope of iJOIN.

Table 8-7: Area throughput gain per scenario

		AT gain deriving from combined CTs	AT gain deriving from network densification	AT gain derived from higher MIMO order in 5G	Overall iJOIN gain
Scenario 1	DL	3.12x	10x	2x	62.4x
	UL	2.6x			52x
	Avg	**2.86x**			**57.2x**
Scenario 2	DL	3.3x	10x	2x	66x
	UL	2.6x			52x
	Avg	**2.95x**			**59x**
Scenario 3	DL	1.3x	10x	2x	26x
	UL	4.83x			96.6x
	Avg	**3.07x**			**61.3x**
Scenario 4	DL	2.4x	10x	2x	48x
	UL	1.26x			25.2x
	Avg	**1.83x**			**36.6x**

Our initial goal for cost-efficiency has been 10% of the baseline architecture. During iJOIN's project time, this metric has been slightly changed to measure the CAPEX compared to a distributed non-centralised deployment to show that with iJOIN's technologies the CAPEX is reduced to 75% of a distributed architecture [37], which implies that a centralised operation of a small-cell network over heterogeneous backhaul offers both improved area-throughput performance and reduced CAPEX. Furthermore, it is worth to underline that the key economic benefit of centralised RAN comes from OPEX, as having a software-based operation is much more cost-efficient. However, the analyses on the corresponding savings were conducted by operators and are highly confidential and cannot be reported here.

Finally, iJOIN has shown that the utilisation of the network can be significantly increased to 90% and higher compared to strongly under-utilised networks today and the initial target of 75% utilisation. This has been achieved with novel cell-association, mobility, and data processing algorithms.

9 Summary and Conclusions

This deliverable provided a comprehensive overview of the final iJOIN architecture. This comprises the definition of a logical, functional, physical, and cloud architecture in order to implement the iJOIN system concept in future evolution of the LTE network. Furthermore, it provides a conclusive description of system design criteria. Based upon these criteria, this report provided a quantitative performance analysis of the iJOIN system concept. In addition, this report described how the iJOIN functional split concept and joint RAN/BH design can be implemented. Based on this report, the following conclusions can be drawn:

- Depending on the actual scenario, the iJOIN system concept achieves area throughput improvements of up to 62x, an energy efficiency of 10% compared to baseline 3GPP LTE R10, a utilisation efficiency of above 90%, and a cost-efficiency in terms of CAPEX of 75% compared to a conventional distributed implementation,
- The iJOIN system concept is applicable to the ETSI NFV architecture and may be implemented on commodity computing hardware while maintaining the real-time requirements imposed by the 3GPP LTE system,
- By means of iJOIN's system design criteria, iJOIN was able to integrate technologies with very different operating regimes, i.e. on PHY layer (short timescale), MAC/RRM layer (medium timescale), and network management layer (long timescale),
- The iJOIN system concept can be integrated transparently into existing 3GPP LTE network in order to allow for a graceful migration towards 5G network based upon iJOIN technologies,
- Both iJOIN key innovations, functional split and joint RAN/BH design, allow for a flexible centralisation in small-cell network deploying heterogeneous backhaul technologies, thereby achieving the performance gains described earlier.

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).

References

[1] iJOIN Project, “D5.1- Revised definition of requirements and preliminary definition of the iJOIN architecture”, November 2013. Available: http://www.ict-ijoin.eu/wp-content/uploads/2014/01/D5.1.pdf

[2] 3GPP, “TR 36.932 V12.0.0; Scenarios and Requirements for Small Cell Enhancements for E-UTRA and E-UTRAN”, Dec. 2012.

[3] iJOIN Project, “D2.1- State-of-the-art of and promising candidates for PHY layer approaches on access and backhaul network”, November 2013. Available: http://www.ict-ijoin.eu/wp-content/uploads/2014/01/D2.1.pdf

[4] iJOIN Project, “D2.2- Definition of PHY layer approaches that are applicable to RANaaS and a holistic design of backhaul and access network”, November 2014. Available: http://www.ict-ijoin.eu/wp-content/uploads/2012/10/D2.2.pdf

[5] 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.

[6] 3GPP TS 36.300 (v10.8.0), “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 10)”, Jul. 2012.

[7] D. Lopez-Perez, X. Chu, and I. Guvenc, “On the Expanded Region of Picocells in Heterogeneous Networks”, *IEEE J. Sel. Top. Sign. Proces.*, vol. 6, no. 3, pp. 281–294, 2012.

[8] iJOIN Project. “D2.3- Final definition and evaluation of PHY layer approaches for RANaaS and joint backhaul/access design”, May 2015.

[9] iJOIN Project, “D3.1- Final report on MAC/RRM state-of-the-art, Requirements, scenarios and interfaces in the iJOIN architecture”, November 2013. Available: http://www.ict-ijoin.eu/wp-content/uploads/2014/01/D3.1.pdf

[10] iJOIN Project. “D3.2- Definition of MAC and RRM approaches for RANaaS and a joint backhaul/access design”, November 2014. Available: http://www.ict-ijoin.eu/wp-content/uploads/2012/10/D3.2.pdf

[11] iJOIN Project. “D3.3- Final definition and evaluation of MAC and RRM approaches for RANaaS and joint backhaul/access design”, May 2015.

[12] iJOIN Project. “D4.1- Report on SotA and requirements for network-layer algorithms and network operation and management”, November 2013. Available: http://www.ict-ijoin.eu/wp-content/uploads/2014/01/D4.1.pdf

[13] iJOIN Project. “D4.2- Network-layer algorithms and network operation and management: candidate technologies specification”, November 2014. Available: http://www.ict-ijoin.eu/wp-content/uploads/2012/10/D4.2.pdf

[14] iJOIN Project, “D4.3- Final definition and evaluation of network-layer algorithms and network operation and management”, May 2015.

[15] iJOIN Project, “- Final definition of Requirements and Scenarios”, November 2014. Available: http://www.ict-ijoin.eu/wp-content/uploads/2012/10/D5.2.pdf

[16] iJOIN Project. “D6.1- Preliminary proof-of-concept results for selected candidate algorithms,” November 2014. Available: http://www.ict-ijoin.eu/wp-content/uploads/2012/10/D6-1.pdf

[17] iJOIN Project, “D6.2- Final proof-of-concept results for selected candidate algorithms,” May 2015.

[18] J. Bartelt, P. Rost, D. Wübben, J. Lessmann, B. Melis, G. Fettweis, “Fronthaul and Backhaul Requirements of Flexible Centralization in Cloud Radio Access Networks,” submitted to IEEE Wireless Communication Magazine, January 2015.

[19] D. Sabella, A. De Domenico, E. Katranaras, M. A. Imran, M. Di Girolamo, U. Salim, M. Lalam, K. Samdanis, and A. Maeder, "Energy Efficiency Benefits of RAN-as-a-Service Concept for a Cloud-Based 5G Mobile Network Infrastructure", Access, IEEE , vol.2, no., pp.1586,1597, 2014. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6990725&punumber%3D6287639

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

[21] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, A. Fehske, "How much energy is needed to run a wireless network?", Wireless Communications, IEEE , vol.18, no.5, pp.40,49, October 2011.

[22] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, I. Godor, "Reducing Energy Consumption in LTE with Cell DTX", Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd , pp.1,5, 15-18 May 2011.

[23] 3GPP TSG RAN, TR 36.927, "Evolved Universal Terrestrial Radio Access (E-UTRA); Potential solutions for energy saving for E-UTRAN (Release 12)", V12.0.0, September 2014.

[24] ETSI, "GS NFV 002 V1.2.1; Network Functions Virtualisation (NFV); Architectural Framework", Dec. 2014.

[25] 3GPP, "TS 36.401 V12.1.0; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Architecture description", Dec. 2014.

[26] ETSI, "GS NFV-SWA 001 V1.1.1; Virtual Network Functions Architecture", Dec. 2014.

[27] NGMN (Next Generation Mobile Networks Alliance), "5G White Paper", NGMN, February 2015.

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

[29] Monica Paolini, "Charting the path to RAN virtualization: C-RAN, fronthaul and HetNets", Senza Fili Consulting, February 2015.

[30] F. Rayal and J. Madden, "Cloud RAN: Enabling NFV in Mobile Networks", XONA Partners and Mobile Experts, February 2015.

[31] NEC Corporation, "iPASOLINK EX: PASOLINK | NEC", [Online]. Available: http://www.nec.com/en/global/prod/nw/pasolink/products/ipasolinkEX.html

[32] U. Dötsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative Analysis of Split Base Station Processing and Determination of Advantageous Architectures for LTE", Bell Labs Technical Journal, vol. 18, no. 1, pp. 105–128, May 2013.

[33] H. Paul, D. Wübben, and P. Rost, "Implementation and Analysis of Forward Error Correction Decoding for Cloud-RAN Systems", IEEE International Conference on Communications 2015 (ICC'15), Workshop on Cloud-Processing in Heterogeneous Mobile Communication Networks (IWCPM), London, England, June 2015.

[34] J. Bartelt and G. Fettweis, " An Improved Decoder for Cloud-Based Mobile Networks under Imperfect Fronthaul", IEEE Global Conference on Communications 2015 (GC'15), Workshop on Wireless Optical Network Convergence in support of Cloud Architectures (WONC), Austin (TX), USA, December 2015.

[35] P. Rost, S. Talarico, and M. Valenti, "The Complexity-Rate Tradeoff of Centralized Radio Access Networks", minor revision submitted to IEEE Transactions on Wireless Communications, March 2015, Available: http://arxiv.org/abs/1503.08585

[36] S. Bhaumik, S.P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, "CloudIQ: A framework for processing base stations in a data center", IEEE MobiCom, Istanbul, Turkey, August 2012.

[37] V. Suryaprakash, P. Rost, and G. Fettweis, "Are heterogeneous cloud-based radio access networks cost-effective?", accepted for publication in IEEE Journal on Selected Areas of Communications, February 2015, Available: http://arxiv.org/abs/1503.03366

[38] C.-L. I, J. Huang, R. Duan, C. Cui, J.X. Jiang, L. Li, “Recent Progress on C-RAN Centralization and Cloudification”, IEEE Access, August 2014.

[39] ONF, "OpenFlow Configuration and Management Protocol OF-CONFIG 1.0", June 2012.

[40] Open Networking Foundation, https://www.opennetworking.org/

[41] 3GPP, TR 23.829, “Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO)”, V10.0.1, October 2011.

[42] Yan Chen, Shunqing Zhang, Shugong Xu, and Geoffrey Ye Li: "Fundamental trade-offs on green wireless networks", Communications Magazine, IEEE , vol.49, no.6, pp.30,37, June 2011.

[43] U. Dötsch, M. Doll, H.P. Mayer, F. Schaich, J. Segel, and P. Sehier, “Quantitative Analysis of Split Base Station Processing and Determination of Advantageous Architectures for LTE,” *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, May 2013.

[44] P. Rost and A. Prasad, “Opportunistic HARQ – Enabler of cloud-RAN over non-ideal backhaul,” IEEE Wireless Communications Letters, June 2014.

[45] P. Rost, C.J. Bernardos, A. De Domenico, M. Di Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, *Cloud Technologies for Flexible 5G Radio Access Networks*, IEEE Communications Magazine, May 2014.