Mobile and wireless communications Enablers for the Twenty-twenty Information Society-II

Deliverable D2.2
Draft Overall 5G RAN Design

Version: v1.0

2016-06-30

The 5G Infrastructure Public Private Partnership

http://www.5g-ppp.eu/

Deliverable/Report D2.2
Draft Overall 5G RAN Design

Grant Agreement Number:	671680
Project Name:	Mobile and wireless communications Enablers for the Twenty-twenty Information Society-II
Project Acronym:	METIS-II
Document Number:	METIS-II/D2.2
Document Title:	Draft Overall 5G RAN Design
Version:	v1.0
Delivery Date:	2016-06-30
Editor(s):	Patrick Marsch, Nokia Bell Labs Ömer Bulakci, Huawei Technologies ERC Icaro Da Silva, Ericsson
Authors:	Paul Arnold, Nico Bayer, Jakob Belschner, Thomas Rosowski, Gerd Zimmermann, **Deutsche Telekom** Mårten Ericson, Icaro Leonardo da Silva, **Ericsson** Ömer Bulakci, Alexandros Kaloxylos, Panagiotis Spapis, **Huawei Technologies ERC** Ahmed M. Ibrahim, Yang Yang, **Intel** Shubhranshu Singh, **ITRI** Haris Celik, **KTH** Jens Gebert, Patrick Marsch, Athul Prasad, Fernando Sanchez Moya, Mikko Säily, **Nokia Bell Labs** Milos Tesanovic, David Gutierrez Estevez, Mehrdad Shariat, **Samsung** Enrico Burrachini, Roberto Fantini, Alessandro Trogolo, **Telecom Italia** Jose F. Monserrat**, Universitat Politecnica de Valencia**
Keywords:	5G, RAN design, requirements, CP/UP design, air interface landscape, harmonization, spectrum, CN/RAN interface, functional design, network slicing
Status:	published
Dissemination level:	Public

Executive Summary

This deliverable provides the consolidated preliminary view of the METIS-II partners on the 5th generation (5G) radio access network (RAN) design at a mid-point of the project.

The overall 5G RAN is envisaged to operate over a **wide range of spectrum bands** comprising of heterogeneous spectrum usage scenarios. More precisely, the 5G air interface (AI) is expected to be composed of multiple so-called AI variants (AIVs), which include evolved legacy technology such as Long Term Evolution Advanced (LTE-A) as well as novel AIVs, which may be tailored to particular services or frequency bands. There is the common view that there should be a **high degree of harmonization among the different AIVs** introduced in 5G, for instance enabling a large-scale reuse of network functions and processing blocks for the sake of reduced implementation complexity and a lean standards specification.

The METIS-II partners further share the view that there should be a **logical split between core network (CN) and RAN**, allowing for an independent evolution of both domains, though it is expected that there will be some **shift of functionality from the CN to the RAN**, for instance related to mobility and paging. It is generally foreseen that **LTE-A evolution and novel AIVs should be integrated on RAN level**, for instance based on user plane aggregation on packet data convergence protocol (PDCP) level, enabling a fast setup of novel radio links and fast switching among these. Such RAN-level integration suggests to also have common CN functions, and a **common interface between CN and RAN** for the different radio technologies.

There are various considerations on how the protocol stack functions could differ in 5G compared to legacy technology. For instance, **functions could be tailored to specific services**, or certain functions may be turned on or disabled for certain services, with various examples provided in this document. There is the consideration to move those functions of the radio link control (RLC) layer which are typically operated in a time-synchronous manner to the medium access control (MAC) layer, while keeping separate logical channels. This way, there is a clear split between asynchronous and synchronous functionalities between the new RLC and MAC, which could be a good function split point for centralized and distributed deployments. Further, there is the common view that **certain network functions could operate on a faster time scale in 5G than in legacy systems**. For instance, traffic steering may not be performed in the form of handover on radio resource control (RRC) level between technologies, as between 3G and 4G, but could be done in a much more agile way and on a faster time scale and on lower layers within the RAN when applied among 5G AIVs.

Beyond general design aspects, the deliverable lists specific functional design considerations developed in METIS-II, such as a **novel RRC state** and related mobility functions enabling device-driven mobility for inactive devices that does not involve CN / RAN signaling, and **RAN-based paging** allowing the tracking of devices on cell-level. Further, detailed considerations on the possible application of **agile resource management (RM) among 5G AIVs** are provided.

Contents

List of Abbreviations and Acronyms

3GPP	3rd Generation Partnership Project
4G	4th Generation of mobile networks
5G PPP	5th Generation Public-Private-Partnership
ACDC	Application specific Congestion control for Data Communication
AI	Air Interface
AIV	AI Variant
AM	Acknowledged Mode
AMC	Adaptive Modulation and Coding
API	Application Programming Interface
ARQ	Automatic Repeat Request
AS	Access Stratum
BLER	BLock Error-Rate
BS	Base Station
BSS	Business Support Systems
C-RAN	Centralized/Cloud-RAN
CAPEX	Capital Expenditure
CN	Core Network
CoMP	Coordinated MultiPoint
CP	Control Plane
CPRI	Common Public Radio Interface
CQI	Channel Quality Indicator
CSI	Channel State Information
CU	Cellular User
D2D	Device-to-Device
DL	Downlink
DMRS	DeModulation Reference Signal
DRX	Discontinuous Reception
E-UTRAN	Evolved-UTRAN
E2E	End-to-End
ECM	EPS Connection Management
eICIC	Enhanced ICIC
EM	Element Managing

EPC	Evolved Packet Core
EPS	Evolved Packet System
ETSI	European Telecommunications Standards Institute
FBMC	Filterbank Multi-Carrier
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FQAM	Frequency Shift Keying and Quadrature Amplitude Modulation
HARQ	Hybrid ARQ
HO	HandOver
HSDPA	High Speed Downlink Packet Access
HW	HardWare
GHz	Giga Hertz
I2I	Indoor to indoor
I/Q	In-phase/Quadrature
ICIC	Inter-Cell Interference Coordination
IE	Information Element
IFFT	Inverse FFT
IMS	IP Multimedia Sub-system
IMT	International Mobile Telecommunications
IMT-2020	IMT for year 2020 and beyond
IoT	Internet of Things
IP	Internet Protocol
ISG	Industry Specification Group
ITU	International Telecommunication Union
ITU-R	ITU – Radiocommunication Sector
KPI	Key Performance Indicator
L2/L3	Layer 2 (MAC + RLC + PDCP) / Layer 3 (RRC)
LAA	License Assisted Access
LSA	Licensed Shared Access
LTE (-A)	Long Term Evolution (-Advanced)

LTE-M	Variant of LTE for M2M communications
MAC	Medium Access Control
MANO	Management and Orchestration
MBB	Mobile Broadband
MBMS	Multimedia Broadcast Multicast System
MBSFN	Multicast-Broadcast Single-Frequency Network
MC	Multi Connectivity
MIMO	Multiple-Input Multiple-Output
mMTC	Massive MTC
MTC	Machine-Type Communications
NAS	Non-Access Stratum
NF	Network Function
NFV	Network Function Virtualization
NGMN	Next Generation Mobile Networks
NLoS	Non-Line-of-Sight
O2I	Outdoor to indoor
O2O	Outdoor to outdoor
OAM	Operations, Administration and Maintenance
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operational Expenditure
OSS	Operations Support System
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDU	Protocol Data Unit
P-GW	Packet Gateway
PHY	Physical layer
QoE	Quality of Experience
QoS	Quality of Service
RACH	Random Access CHannel
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RF	Radio Frequency
RLC	Radio Link Control

RLF	Radio Link Failure
RM	Resource Management
RNC	Radio Network Controller
RRC	Radio resource control
RRM	Radio RM
RRU	Remote Radio Unit
RS	Reference Signal
S-GW	Serving Gateway
SAP	Service Access Point
SDN	Software Defined Networking
SDR	Software Defined Radio
SDU	Service Data Unit
SINR	Signal-to-Interference and Noise Ratio
SLA	Service Level Agreement
SN	Sequence Number
SON	Self Organizing Networks
SRS	Sounding Reference Signals
SW	SoftWare
TB	Transport Block
TeC	Technology Component
TCP	Transmission Control Protocol
TDD	Time-Division Duplex
TTI	Transmit Time Interval
UC	Use Case
UDN	Ultra-dense Network
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UM	Unacknowledged Mode
uMTC	Ultra-reliable MTC
UMTS	Universal Mobile Telecommunications System
UP	User Plane
UTRAN	Universal Terrestrial RAN
V2X	Vehicle-to-Anything
VNF	Virtual Network Function
VoLTE	Voice over LTE
WG	Working Group

WLAN	Wireless Local Area Network
WP	Work Package
WRC-15	World Radiocommunication Conference in 2015
WRC-19	World Radiocommunication Conference in 2019
xMBB	Extreme Mobile Broadband
xHaul	Backhaul / Midhaul / Fronthaul

1 Introduction

1.1 Motivation and Scope of this Deliverable

The main objective of the METIS-II project is to develop an overall 5G radio access network (RAN) design according to “technology level 2”, and obtain consensus among key players in the field on the key RAN design paradigms therein. To achieve this objective, METIS-II builds upon technology components (TeCs) that have already been developed in earlier projects (e.g. METIS [METIS] or 5G NOW [5GNOW]), or which are being developed in METIS-II as such, complements these with any enablers that are required, and develops comprehensive overall functionality frameworks for the 5G RAN. This overall process, which is being pursued in individual work packages (WPs) in METIS-II as well as in the overall RAN design, is depicted in Figure 1-1.

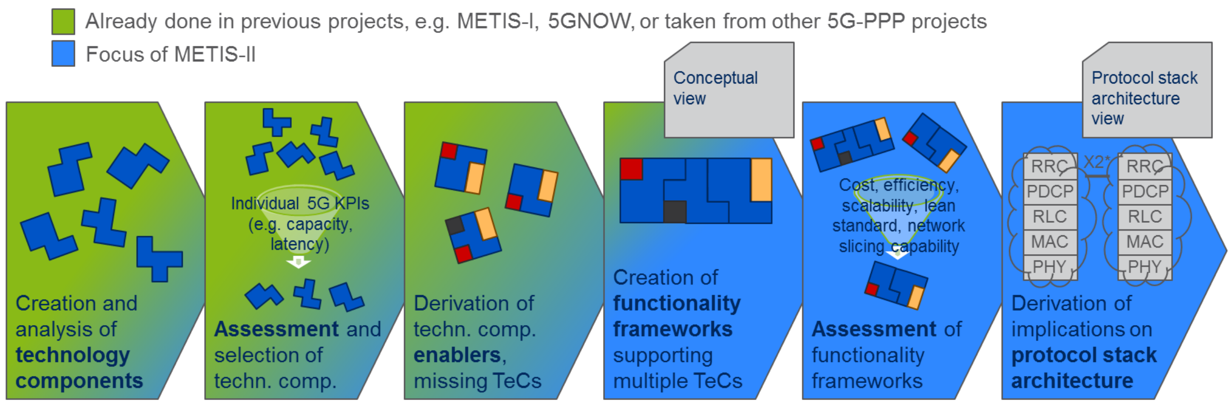


Figure 1-1. 5G RAN design process pursued in the different technical WPs in METIS-II.

This deliverable captures the status of the METIS-II 5G RAN design at a mid-point in the project. In particular, it highlights the consensus that has already been found on key RAN design questions that were posed at the beginning of the project, and describes the current view on the overall 5G RAN architecture and its functional design. It has to be noted that different aspects of the 5G RAN design have obtained a different level of maturity so far, as also indicated in Figure 1-1. For some aspects, in particular where a rich set of TeCs was already available before, the work has focused on the integration of these (as for instance the works related to air interface (AI) design, see Section 4.2). In other fields, however, the METIS-II focus has so far been on the development of additional and missing TeCs (as for instance aspects related to the functional design of the 5G RAN, see Chapter 6).

It should further be noted that this deliverable only provides a compact high-level overview on the METIS-II 5G RAN design, summarizing various other deliverables that have recently published. The relation of these deliverables is depicted in Figure 1-2. The reader is highly encouraged to refer to the stated other deliverables under [METIS].

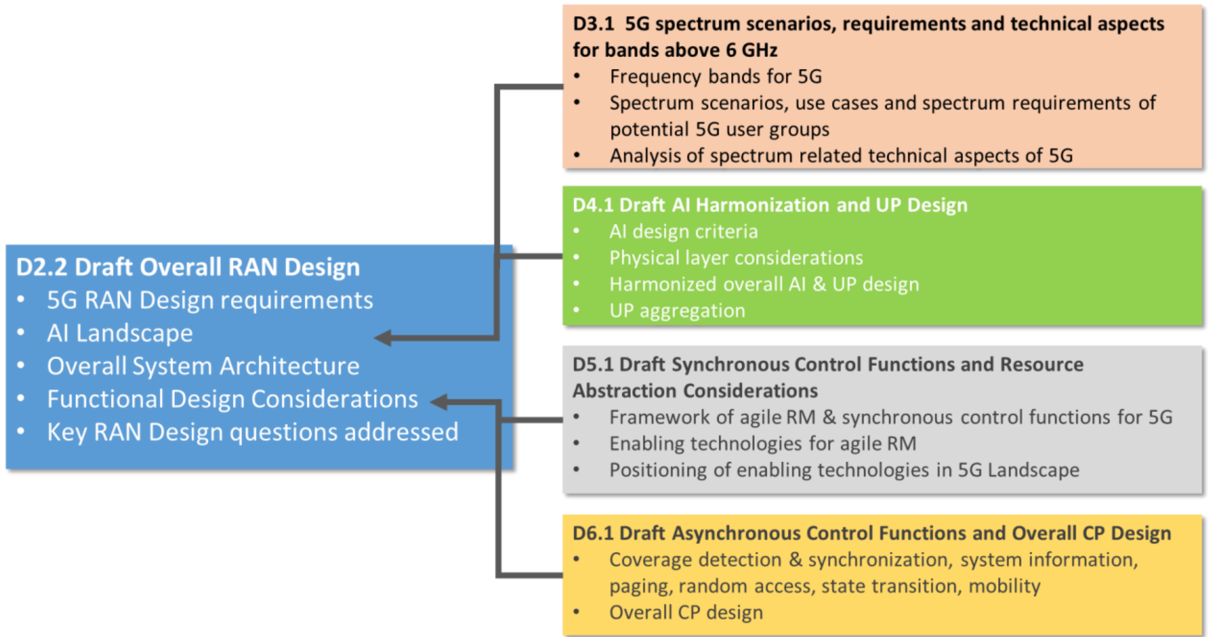


Figure 1-2. Relation of this deliverable to other deliverables with more technical depth.

Please note that this current deliverable, and the related in-depth works captured in Figure 1-2, correspond to the preliminary output of METIS-II at a mid-point in the project. Significantly more results and details on the 5G RAN design will be disseminated throughout the remainder of the project, as shown in form of the project timeline in Figure 1-3. We would in particular like to emphasize that

- simulation results backing up various of the concepts discussed in this deliverable will be provided in the form of R2.3 in October 2016, and in the form of D2.3 in February 2017;
- an interactive evaluation and visualization of key concepts described in this document will be provided in towards the end of 2016, and captured in D7.2;
- the final METIS-II 5G RAN design will be disseminated through various deliverables in the time frame April-June 2017.

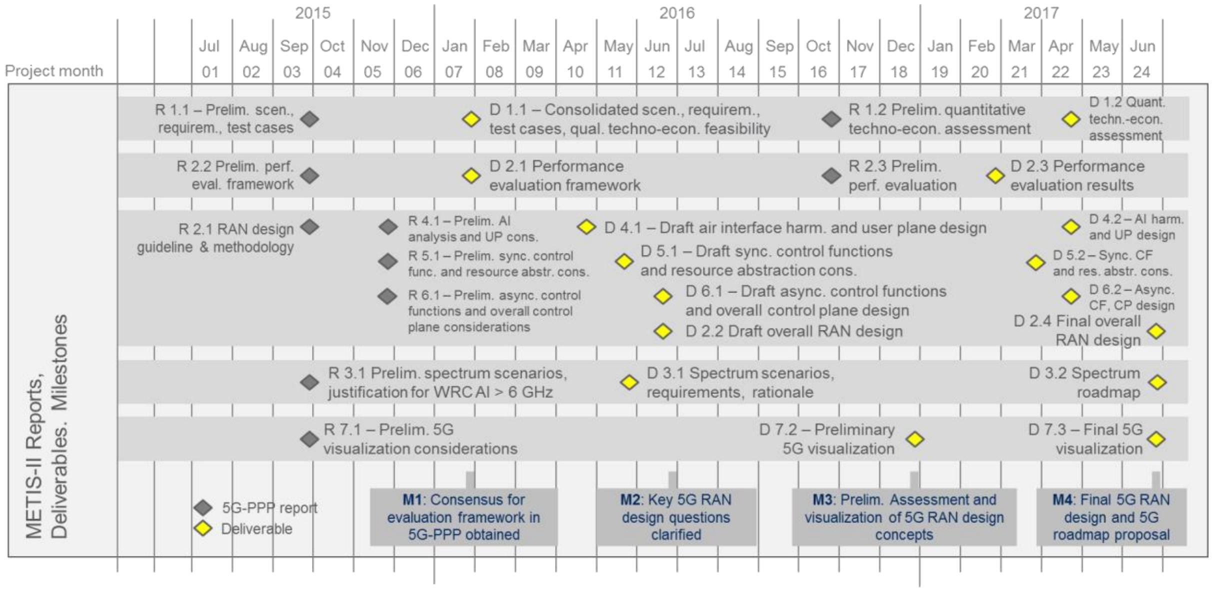


Figure 1-3. Project timeline and key milestones of METIS-II.

1.2 Structure of this Deliverable

This deliverable is structured as follows: **Chapter 2** provides a compact high-level overview on the 5G service landscape envisioned by METIS-II, and the key innovation pillars seen as most relevant to address this. **Chapter 3** then lists the RAN design requirements as identified by the METIS-II project, including also specific requirements related to the support of a very diverse service landscape, network slicing, and an overall integrated AI in 5G. **Chapter 4** provides an overview on 5G spectrum aspects and the current AI considerations of METIS-II, e.g. showing how different AI variants (AIVs) could possibly be integrated to jointly cover the needs for all services and bands in 5G. It further elaborates on the question how harmonized the design of different AIVs could be, for the sake of a lean standard and reduced implementation complexity. **Chapter 5** then provides a summary on the overall 5G RAN architecture view of METIS-II, covering the envisioned split between core network (CN) and RAN, and related logical CN/RAN and also intra-RAN interfaces. The chapter further ventures into the protocol stack architecture in 5G and the likely changes to protocol stack layers that should be introduced in the 5G time frame, as well as envisioned physical deployment options in 5G and the likely mapping of logical to physical architecture. The chapter also covers how the 5G RAN architecture is expected to support network slicing, and network management and orchestration aspects. Beyond the overall RAN architecture, **Chapter 6** then ventures into various detailed functional design considerations of METIS-II, for instance related to traffic steering and RM in 5G, or initial access and mobility. **Chapter 7** provides a summary on the METIS-II status of answering the key 5G RAN design questions that were posed at the beginning of the project, and **Chapter 8** concludes this deliverable.

2 The METIS-II Vision on 5G

2.1 Envisioned 5G Service Landscape

There is already a wide consensus on the 5G service landscape, and in particular on the view that 5G will not only be a "business-as-usual" evolution of 4G mobile networks, with new spectrum bands, higher spectral efficiencies and higher peak throughputs, but will also target new services and new business models. These latter are to be developed in close collaboration with vertical industries and imply new requirements and new ways of thinking, building and managing the network. The analysis of the needs and requirements of these verticals has lead the METIS project [MET15-D15], and forums such as Next Generation Mobile Networks (NGMN) [NGM15-WP], the International Telecommunication Union (ITU)-R [ITU15] and the 3rd Generation Partnership Project (3GPP) [3GPP16-38913] to consider the following three main 5G service types:

- **Extreme Mobile BroadBand (xMBB)**, often referred to as enhanced mobile broadband (eMBB), requiring both extremely high data rates and low-latency communication in some areas, and reliable broadband access over large coverage areas.
- **Massive Machine-Type Communications (mMTC)**, requiring wireless connectivity for up to tens of billions of network-enabled devices worldwide. Here, scalable connectivity for an increasing number of devices per cell, wide area coverage and deep indoor penetration are key priorities.
- **Ultra-reliable Machine-Type Communications (uMTC)**, often referred to as ultra-reliable and low-latency communications (URLLC), e.g. related to vehicle to anything (V2X) communication [3GPP16-38913], industrial control applications, Smart Grid etc.

It goes without saying that considering each service type separately and building a 5G network accordingly, we would likely end up with very different radio access network (RAN) designs and architectures. However, only a common RAN that accommodates all three service types will likely be an economically and environmentally sustainable solution. For this reason, the RAN design of METIS-II is performed specifically towards a set of 5G use cases that typically combine multiple service types. More precisely, the project has performed an analysis of the 5G use cases considered by various stakeholders, classified them into families considering the special characteristics of these (e.g., services covered, mobility, and/or number of users, infrastructure, etc.), and has chosen five use cases that are seen as most representative of these different families [MII16-D11]. These use cases, with their key requirements, are depicted in Figure 2-1. As part of the 5th Generation Public-Private-Partnership (5G PPP) cross-project activities performed by METIS-II, the service and use case considerations listed above have been discussed with other projects in 5G PPP, and the joint view of various projects on use cases and models has been consolidated and captured in a joint 5G PPP document [5GPPP16-UCM].

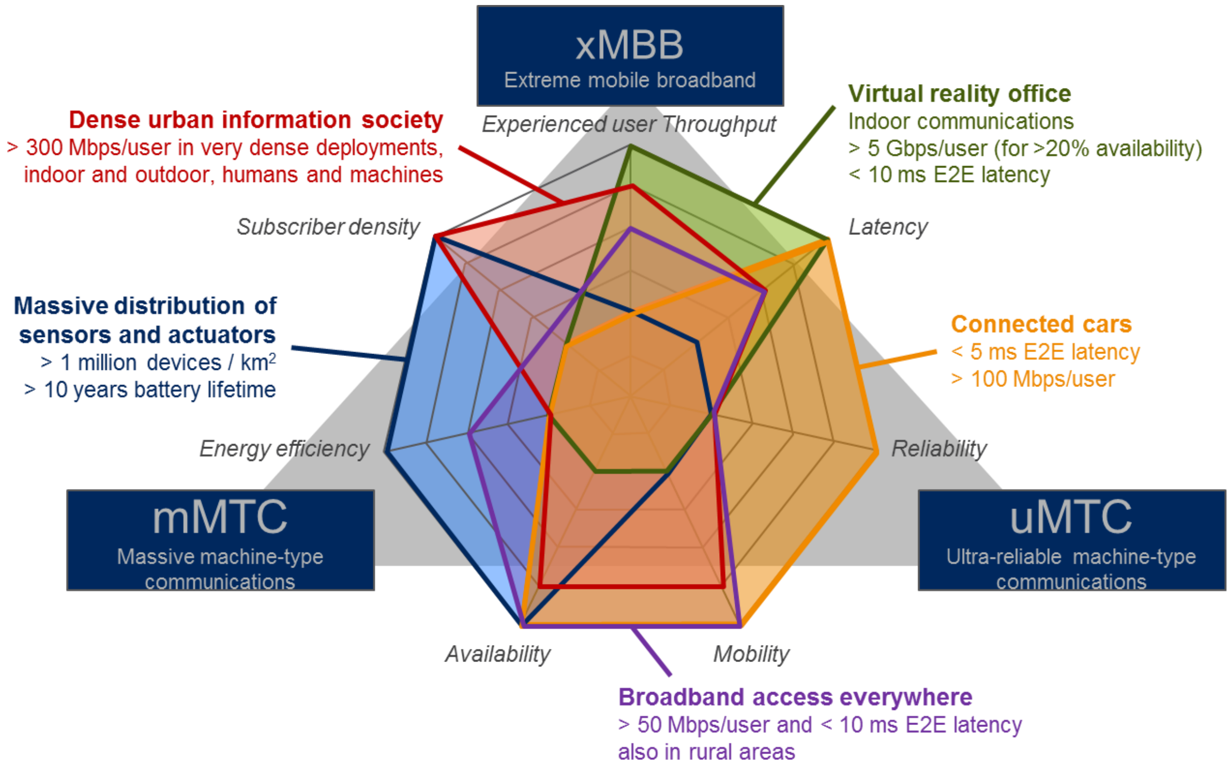


Figure 2-1. Considered main 5G service types and representative use cases [MII16-D11].

The main key performance indicators (KPIs) related to the identified services and use cases are:

1. Experienced end-user throughput, targeting values of 300 Mbps in DL and 50 Mbps in UL in dense urban environment and up to 1 Gbps for virtual reality offices,
2. reliability, achieving values of up to 99.999% for end-to-end latencies < 5 ms, with the main application area being traffic efficiency and safety in the V2X context,
3. a massive number of devices, with more than 1 million devices per km^2,
4. availability and retainability,
5. high mobility, with device speeds up to 250 km/h, and
6. energy consumption (energy efficiency).

2.2 Key Innovation Pillars needed for 5G

In order to enable the previously stated services and their requirements, METIS-II has already envisioned in the project proposal that in particular the following key fields require substantial innovation, and is consequently putting a strong emphasis on these in its work, as shown in Figure 2-2:

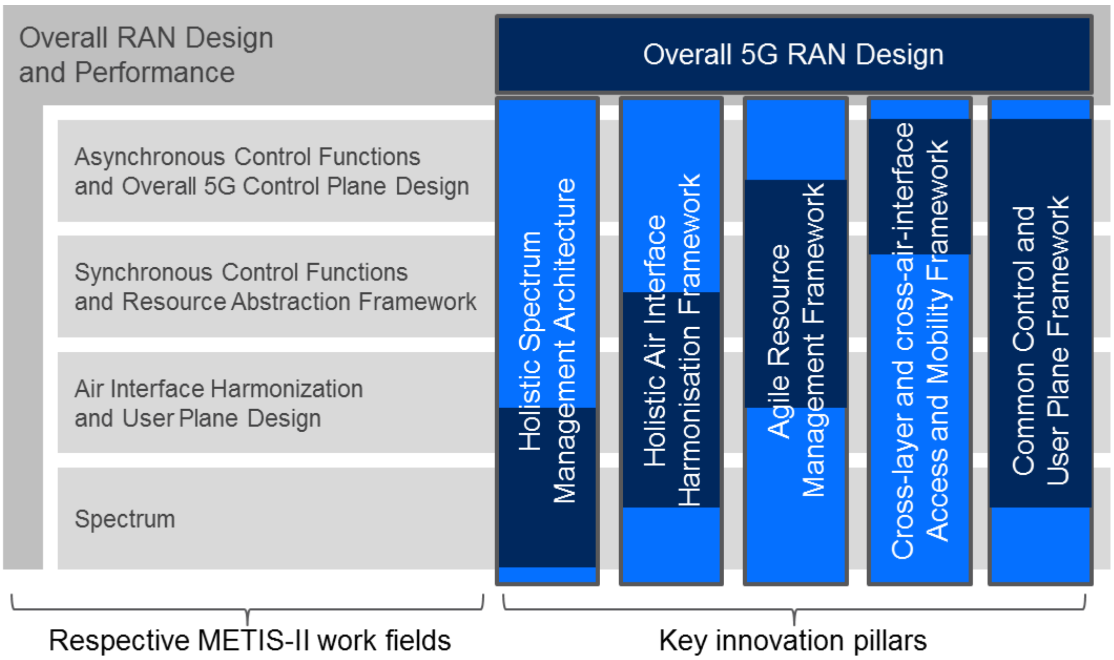


Figure 2-2. Work fields and key innovation pillars addressed in METIS-II.

- A **Holistic spectrum management architecture** is required for flexible spectrum management and multi-operator collaboration in 5G. This should in particular address the integration of spectrum bands above 6 GHz, as the usage of these bands is novel in cellular communications and subject to various challenges as highlighted in Section 4.1. Further, the architecture should address “new 5G user groups” like vertical industries, and the efficient use of different spectrum authorization modes and access schemes. The progress of METIS-II in this respect is captured in Section 4.1.

- A **Holistic air interface (AI) harmonization framework** is required to ensure that similar protocol functionalities in the different protocol layers of AIVs to be used in 5G are harmonized to a strong extent, including both legacy and new AIVs, with the aim to keep device and infrastructure complexity tractable and ensure a lean standard. The latest METIS-II considerations on this topic are captured in Section 4.2.

- An **Agile Resource Management (RM) framework** aims to provide holistic RM solutions and AI abstraction models that consider and exploit the novel aspects of 5G systems, such as, very diverse service requirements, existence of multiple AIVs in the overall AI, dynamic topologies (e.g., based on moving networks), and novel communication modes (e.g., device-to-device, D2D). In the scope of agile RM, METIS-II extends the notion of a resource beyond conventional radio RM (RRM) and aims to attain the optimum mapping of 5G services to any available resources when and where needed within this extended realm of resources. Accordingly, to the framework comprises the re-design of functions, e.g. related to interference management and traffic steering, as well as new functions such as RM for network slices. This is covered in detail in Sections 6.1 and 6.2.

- With the **Cross-layer and cross-air-interface system access and mobility framework**, the cross-layer optimization concept is extended to another dimension, enabling a higher interaction of functionalities from the multiple AIVs in order to improve the overall resource usage. New procedures have to be designed in such a way that functionalities of one AIV could be used in a flexible way by another AIV, where this coordination is facilitated by a control plane framework common for all AIVs. This is covered in detail in Section 6.3.

- A **Common control and user plane framework** for multiple 5G AIV, including the evolution of legacy standards such as Long Term Evolution-Advanced (LTE-A), is seen as essential to enable fast link switching, control/user plane diversity, throughput aggregation, etc. in order to fulfil the diverse and stringent 5G requirements. This is covered in detail in the context of describing the envisioned overall 5G RAN architecture in Chapter 5.

3 Key 5G RAN Design Requirements

Going beyond the aforementioned key innovation fields that were already listed before the project start, METIS-II has during its runtime identified the following general architectural or functional requirements according to which the 5G RAN should be designed in order to meet the diverse service requirements stated in Section 2.1:

- The **5G RAN should be highly scalable** in terms of throughput, the number of devices, the number of connections etc. To enable this, it should be able to **handle and scale user plane (UP) and control plane (CP) independently**. Further considerations on the support of diverse service requirements are listed in Section 3.1.

- **The 5G RAN should support the Network Slicing[1] vision from NGMN** [NGM15-WP], aiming to address the deployment of multiple logical networks as independent business operations on a common physical infrastructure. The implication of Network Slicing on the RAN design is a METIS-II research topic by itself and is also elaborated in more detail in Section 3.1 and later in Section 5.6.

- One enabler for the system to handle the diverse service requirements stated before is that **the overall network (both RAN and CN) should be software-configurable**. This means, for instance, that it is configurable which sets of logical and physical entities are to be traversed by CP and UP packets.

- The 5G RAN must be **designed to operate in a wide spectrum range** with a diverse range of characteristics such as bandwidths and propagation conditions, as discussed in Section 4.1. For higher frequency bands such as mmWave bands, beamforming will become essential, for instance in the form of massive multiple input multiple output (MIMO) technology. Therefore, the RAN should support procedures that rely on beamforming in an efficient way.

- The 5G RAN should enable a tight interworking between **LTE-A evolution and novel 5G radio technology on RAN level.** It has to be noted that this does not imply any particular relationship between LTE-A evolution and novel 5G radio, e.g. both technologies may for instance act as mobility anchor for the other, or be operated stand-alone, as described in more detail in Section 5.4.

[1] A “network slice” supports the communication service of a particular connection type with a specific way of handling the CP and UP for this service throughout core network (CN) and RAN, and is seen from a customer perspective as a separated logical network [NGM15-WP, MII16-WP].

- The **5G RAN should natively and efficiently support multi-connectivity,** i.e. the case when the UE is connected to more than one node (inter-node, i.e. not co-located) and / or more than one AI (which may be co-located or not). This is detailed in Section 5.4

- **It should further natively support network-controlled D2D** (i.e. point-to-point, multi-cast and broadcast), and the option that some **5G devices could flexibly act as if they were infrastructure nodes**, one example being self-backhauled, possibly nomadic access nodes. See Section 6.1.5 for details.

- The 5G RAN should be designed such that it can **maximally leverage from centralized processing of radio layers**, **but also operate well in the case of distributed base stations** with imperfect backhaul / midhaul / fronthaul (xHaul) infrastructure, with soft degradation of performance as a function of xHaul quality. More precisely, METIS-II has defined four physical architecture deployment scenarios [MII15-R21], including also a wireless self-backhauling scenario, which should all be supported by any 5G RAN design concepts. Please find more details in Section 5.5.

- The 5G RAN design must be **energy efficient**. For the aim of assessing the energy efficiency of the design, energy-related KPIs have to be assessed following adequate methodologies. A description of energy efficiency KPI and its associated assessment methodology are given in [MII16-D21].

- **The 5G RAN design must be future proof**, i.e. it should enable an efficient introduction of new features and services (e.g., by minimizing the spreading of signals over radio resources and facilitating the introduction of new physical channels) and guarantee backward-compatibility of devices in future releases.

3.1 Design Requirements specifically related to diverse Services and Network Slicing

Beside the aforementioned design requirements, the envisioned **sets of services and their diverse and partially conflicting requirements** will likely pose the following further requirements on the 5G RAN design:

- **Traffic differentiation**: The RAN should support more sophisticated mechanisms for traffic differentiation than legacy systems in order to be able to treat heterogeneous services differently and fulfill more stringent QoS requirements. Potential solutions are described in Section 6.2.

- **Resource reuse**: 5G networks should support a strong reuse of resources (e.g., radio, functional, and infrastructure resources; see *the extended notion of a resource* in [MET16-D51]) to enable an economically viable solution for emerging 5G services.

An efficient joint utilization of infrastructure resources by multiple services and differentiated service treatment also prepare the grounds for Network Slicing in 5G. Beyond these aspects, some additional requirements have been identified that are specific of Network Slicing:

- **Slice-aware RAN**: Slices (or some abstraction thereof, such as particular groups of flows or bearers) should be visible to the RAN CP to enable a treatment related to joint KPIs concerning all services within a slice or across slices. For example, all services within one slice may jointly occupy only a certain extent of some resources (e.g., radio resources and functional resources, as stated above), while other resources (e.g., hardware, HW, and software, SW, platforms) may be shared between slices.
- **Slice protection**: The RAN should support slice isolation, e.g., by providing related slice protection mechanisms so that events within one slice, such as congestion, do not have a negative impact on another slice.
- **Slice management and setup:** The RAN should support efficient management mechanisms, e.g., to efficiently setup and operate slices.
- **Slice-specific network management**: The RAN should allow offering slice-specific network management functions as a service.

3.2 Design Requirements specifically related to Air Interface Integration in 5G

Due to the need of the 5G RAN to support various services and frequency bands, it is clear that there cannot be a one-size-fits-all AI in 5G, e.g. in the sense of a single PHY numerology for 5G, or the exactly same protocol stack instantiation for all services. Instead, there is clear consensus that the overall 5G AI will consist of multiple different AI solutions – in METIS-II termed AIVs – that will be integrated with each other and with evolved LTE-A, see Section 4.2. In this respect, further 5G RAN design requirements related specifically to AIV integration are:

- The 5G RAN should be designed in such a way that a **lean specification** is possible, i.e. that a limited number of specification documents is used to cover multiple services or multiple frequency bands, for instance by using parameterization to allow tailoring of certain concepts, functionalities etc. to different services or bands.
- The **functionalities tailored for different services, bands etc. should be harmonized to the largest extent possible** without sacrificing the performance for individual services, bands etc., to enable maximum SW and HW reuse on network and device side and hence reduced implementation complexity, and to enable user plane aggregation or control plane integration among multiple AIVs. See Section 4.2 for details.

4 Air Interface Landscape envisioned for 5G

The overall 5G RAN is envisaged to operate over a wide range of spectrum bands comprising of heterogeneous spectrum usage scenarios, in order to cope with the demand for higher performance and capacity, but also to provide higher reliability and lower latency. In this chapter, we initially summarize in Section 4.1 the outcome of the spectrum-related work conducted so far in METIS-II, and the recommendations derived from that, and then venture in Section 4.2 into the detailed AI design considerations in METIS-II.

4.1 5G Spectrum Scenarios, Requirements and Aspects for Bands above 6 GHz

4.1.1 Spectrum Scenarios and Requirements for 5G

Radio spectrum usage can generally be authorized in two ways: Individual Authorization (Licensed) and General Authorization (Licence Exempt / Unlicensed). Authorization modes recognized as relevant for wireless communications are *Primary user mode*, *LSA (Licensed Shared Access) mode* and *Unlicensed mode* [MET14-D53]. Furthermore, five basic spectrum usage scenarios can be identified for these authorization modes: dedicated licensed spectrum, limited spectrum pool, mutual renting, vertical sharing and unlicensed horizontal sharing [MII16-D11]. The LAA (License Assisted Access) approach considered for LTE-A is a combination of "dedicated licensed spectrum" and "unlicensed horizontal sharing" by using carrier aggregation [3GPP16-36889].

Frequencies below 6 GHz are likely most suitable to support mMTC services where coverage is most important, whereas spectrum above 6 GHz is essential to provide the massive capacity demanded by xMBB applications. An exclusive use of spectrum should remain the main and preferred solution, while a shared use of spectrum may be a complement to increase spectrum availability [MII16-WP]. The requirements of the METIS-II 5G use cases (see Section 2.1) concerning spectrum can be broadly categorized into three main groups [MII16-D31]:

- **Capacity** to cope with high traffic per area: this can be addressed through higher spectral efficiency, higher site density, but most importantly in this context by a **large amount of preferably contiguous spectrum**.

- **Coverage** to ensure the availability of 5G everywhere: this could partially be addressed through, e.g., massive deployments and massive MIMO, but the most efficient approach is to use **lower frequency bands**.
- **Reliability** to fulfil the demands of critical services, requiring stable and predictable operation conditions: this can for instance be increased through diversity in time, frequency and space (e.g. through multi-connectivity), and is greatly facilitated by having **dedicated spectrum**.

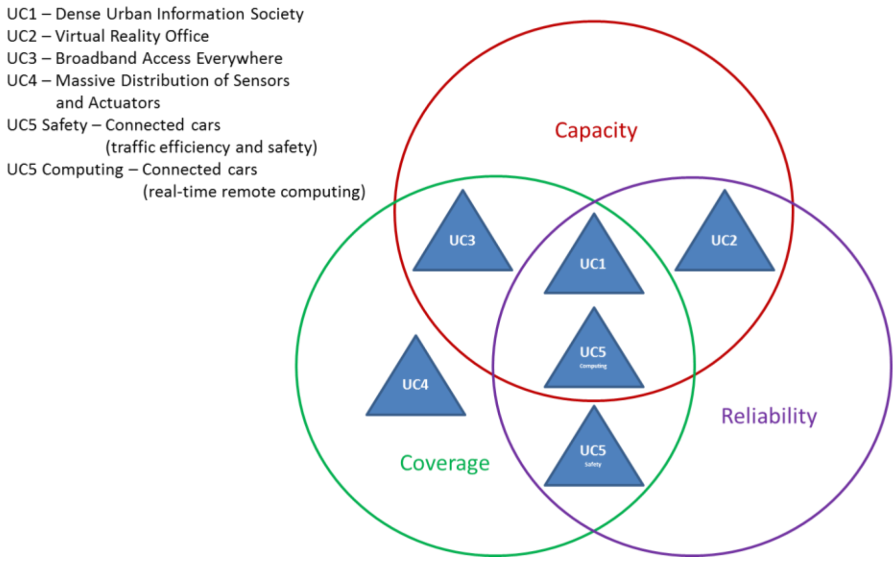


Figure 4-1: Relation between METIS-II 5G use cases and the three spectrum requirement categories.

The relationship between the METIS-II 5G use cases and the three categories defined above is illustrated in Figure 4-1. It is obvious that a combination of different suitable frequency bands is necessary to cope with the use case requirements.

4.1.2 Rationale for 5G Bands above 6 GHz

In [MII15-R31], it is demonstrated that the three basic means to increase wireless network capacity, namely access point density, spectrum efficiency and spectrum bandwidth, are exchangeable to some extent in conventional macro-cell environments. However, densification of access points becomes progressively inefficient in super-dense environments, so that additional spectrum becomes the most effective solution for providing high capacity in such cases. In [MII15-R31] it is also shown that contiguous spectrum offers advantages over multiple fragmented frequency bands with regard to device complexity, signaling overhead, guard bands and interference. Therefore, additional wide contiguous frequency bands are needed to fulfill 5G capacity requirements. WRC-15 has agreed that ITU-R will conduct sharing and compatibility studies for a number of frequency bands between 24.25 GHz and 86 GHz (see Figure 4-2) in time for WRC-19. Some of these frequency bands enable wide contiguous bandwidths, which would allow coping with the requirements of high bandwidth demanding applications.

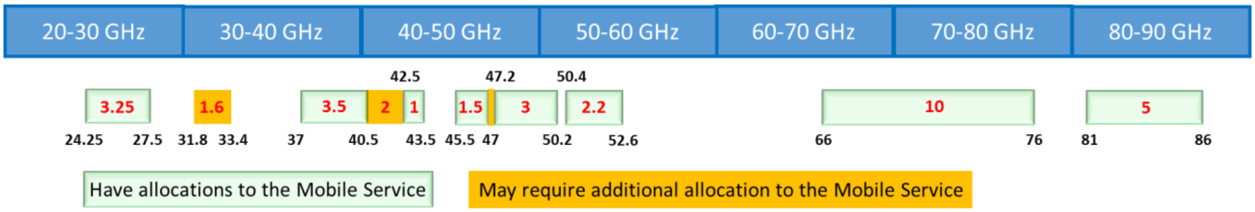


Figure 4-2: Frequency bands to be studied in ITU-R for IMT-2020 until WRC-19.

4.1.3 Coverage aspects for bands above 6 GHz

Assessment results of a 5G system performance evaluation – based on a simple link budget calculation for frequencies up to 100 GHz – indicate that the higher propagation losses with increasing carrier frequencies might be compensated to some extent provided that larger channel bandwidths are available than for lower carrier frequencies, or by implementation of advanced antenna systems. Nevertheless, for the outdoor to indoor (O2I) scenario, the more challenging radio propagation conditions impose more restrictions for the cell size. Depending on propagation conditions and equipment deployed, frequencies up to around 30 GHz are in particular suitable since all three considered stationary scenarios, i.e., O2I, non-light of sight (NLoS) outdoor to outdoor (O2O), and indoor to indoor (I2I), are feasible. With carrier frequencies in the range of 30-60 GHz, the O2O and I2I scenarios appear to be feasible with the considered distance assumptions if advanced beamforming technologies with high antenna gain are applied. With carrier frequencies above 60 GHz, dedicated indoor services could still be feasible, noting that at those frequencies there is also the possibility of obtaining very large contiguous channel bandwidths [MII16-D31]. The coverage feasibility over frequency ranges for different deployment scenarios is illustrated in Figure 4-3.

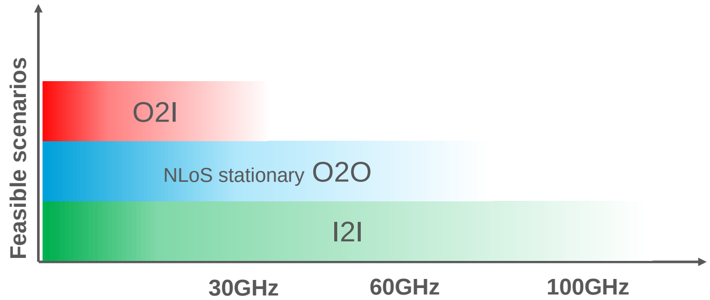


Figure 4-3. Indicative coverage feasibility of different deployment scenarios in different frequency ranges.

4.2 Air Interface Design Considerations

METIS-II considers the overall 5G AI[2] to be comprised of multiple so-called AIVs[3], which may for instance be characterized by tailored numerology and/or features for certain frequency ranges (see previous section), services, or cell types etc.

More precisely, METIS-II envisions that beside the evolution of LTE-A, novel AIVs are needed to fulfill all the performance requirements of the envisioned new use cases. These include some extreme low latency use cases, ultra-reliable transmission and xMBB requiring additional capacity that is only available at very high frequencies, as well as mMTC with extremely densely distributed sensors and very long battery life requirements. Designing an adaptable and flexible 5G AI, which will tackle these use cases while offering native multi-service support, is one of the key challenges in designing a 5G RAN, with far-reaching impact on overall system design.

Further, a key question is how the different AIVs, including LTE-A evolution, can be integrated into one overall 5G AI, such that this design maximally benefits from the wide landscape of bands, cell types etc., and such that both the complexity of the standard and that of the implementation are minimized, while the performance of individual technologies is not sacrificed. To this end, METIS-II has drawn up an evaluation framework for 5G AI candidates, with one key focus on the extent of harmonization across underpinning components in overall AI considerations. The extent of harmonization can be assessed through a set of criteria such as the utilization of radio resources, implementation complexity, standardization effort, forward compatibility, and interaction with legacy systems [MII16-D41]. Additional evaluation criteria include UP-related design principles and requirements posed from CP considerations. The combined evaluation criteria result from wide consensus reached within METIS-II and are well-aligned with 3GPP while offering a long-term, integrated system view.

Details on the METIS-II AI design considerations are provided in [MII16-D41]. In this section, we briefly summarize the key design principles that have been followed, and the different proposals for the overall 5G AI that are currently being investigated.

4.2.1 5G AI Evaluation Criteria and Design Principles

The METIS-II AI candidate selection is performed according to AI evaluation criteria classified into the following four categories:

[2] An AI is here defined as the RAN protocol stack (i.e. PHY / MAC / RLC / PDCP / RRC or 5G equivalents, or subset thereof) and all related functionalities describing the interaction between infrastructure and device and covering **all** services, bands, cell types etc. that are expected to characterize the overall 5G system.

[3] An AIV is defined in the same way as an air interface, but covers only a **subset** of services, bands, cell types expected to characterize the overall system.

- The suitability of an AI proposal to meet the overall 5G KPIs and directly related UP design requirements;
- Additional UP-related AI design principles, which are detailed below;
- Requirements posed from CP considerations on the design of AIs;
- The extent of harmonization across AIVs in overall AI considerations.

The above-mentioned UP-related AI design principles have been derived from the general architectural and functional requirements stated in Chapter 3. These ensure the required flexibility in achieving 5G KPIs which LTE-A and its evolution cannot fulfil in their entirety, and are here described in more detail:

- Flexibility by design: 5G AI needs to be adaptable and flexible in order to provide the required flexibility for multi-service support and non-traditional applications. A single but sufficiently wide harmonized AI would allow this flexibility. More specifically, the extent of harmonization already mentioned is an important METIS-II 5G AI design KPI to achieve this flexibility by design.
- 5G AI should be forward-compatible: This is needed to ensure future-proofness for upcoming variants of existing 5G services as well as potential new services not necessarily in the xMBB, uMTC or mMTC categories. Such a future-proof design needs to allow the introduction of new physical channels.
- 5G AI should offer easy interworking with evolution of LTE-A: It is assumed in METIS-II that the 5G RAN should allow to integrate LTE-A evolution and novel 5G AIVs. The exact mechanics of this interworking are under study in METIS-II.
- The design of the 5G AI should minimize signaling overhead and unnecessary transmissions.
- The 5G AI design should take into account the latest information on bands available (or to be made available shortly) to mobile: 5G systems will operate across a wide range of mmWave and cmWave frequencies. The 5G AI design should, therefore, consider a beam-centric approach, i.e., control and user plane signaling should be designed having in mind that these will often be transmitted in beams.
- The 5G AI design should take into account terminal complexity. The extent of harmonization again plays an important role here, since the implementation of one widely harmonized AI is expected to decrease terminal complexity compared to the implementation of its AI components in a non-harmonized way.
- 5G AI design should enable Application Program Interfaces (APIs) to higher layers, on both device and network sides, so as to facilitate the implementation of network slicing.

Beyond harmonization, METIS-II investigates to which extent UP instances related to different bands can be logically aggregated on certain layers, and beyond which layer there would be a single CP instance. Different AI designs may offer different support of such aggregation and integration features, which also needs to be considered.

4.2.2 Overall 5G AI Proposals under Investigation

A single AI framework is the main goal of METIS-II, but this does not mean a “one-size-fits-all” solution. Different use cases will likely need some level of differentiation in the AI. How to achieve this differentiation is another key question. Additionally, current proposals for overall AI landscapes differ in the technology components they are comprised of, and in the type and their inherent extent of harmonization. As each METIS-II proposal currently under study is a single framework comprised of multiple AIVs selected to fulfill the performance of the different use cases and scenarios, a unified way of describing the 5G AI design proposals using a 5G service / frequency mapping is used, as depicted in Figure 4-4. In this abstract example, it is shown how the overall space of main service types and frequency bands may possibly be covered by one (left side of the figure) or multiple (right side of the figure) waveform families. For the more detailed overall AI considerations listed in the sequel, the same kind of mapping will be used to illustrate how different AIVs are jointly envisioned to cover the overall service and band space.

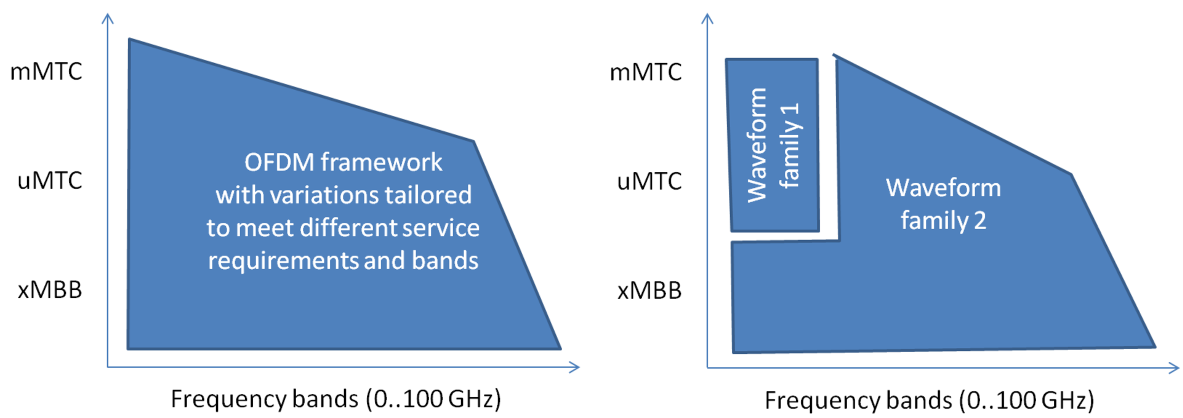


Figure 4-4. Description of 5G AI proposals through service / frequency mapping.

Existing 5G AI proposals differ in the technology components they are comprised of, and in the underlying type and extent of harmonization. The following are the different AI proposals under study in METIS-II:

1. A harmonized layer 1-3 solution based on cyclic prefix orthogonal frequency division multiplex (OFDM) for sub-1 GHz to 100 GHz carrier: scalable Cyclic Prefix-OFDM based solution with a harmonized physical layer (PHY) / medium access control (MAC) / packet data convergence protocol (PDCP) and contention-based access support.
2. Harmonized Cyclic Prefix-OFDM for multiple bands with enhancements for multi-service support: Cyclic Prefix-OFDM based solution with scaling/flexible numerology, utilizing advanced single-carrier (SC) frequency division multiple access (FDMA) and allowing flexible time division duplex (TDD).

3. Multi-service support with universal filtered (UF) OFDM: relies on increased waveform (WF) configurability and reconfiguration (with possible use of cyclic prefix and zero postfix) and provides strong frequency and time confinement.
4. A qualitative, feature-driven AIV design for the 5G landscape: tailoring AIVs to specific service types and harmonizing them starting at the PHY layer.
5. AI design based on pulse-shaped OFDM (P-OFDM): gives flexible numerology and frame structure, uses adaptive pulse shape filter, supports asynchronous multiple access and results in improved resilience against Doppler.
6. Quadrature amplitude modulation (QAM) – filter-bank multi-carrier (FBMC) and OFDM harmonized solution: multi-WF solution (QAM-FBMC for lower frequencies and Cyclic Prefix -OFDM for higher frequencies), support for asynchronous scenarios in lower bands, beamforming solutions for higher bands.
7. OFDM based solution with flexible numerology and frame structure: OFDM-based solution with flexible frame structure, supporting frequency keying and quadrature amplitude modulation (FQAM) in lower bands for enhanced cell-edge performance, beamforming solutions for higher bands.
8. Multi-AIV (offset quadrature amplitude modulation (OQAM)-FBMC, Cyclic Prefix-OFDM) harmonization aspects for above PHY layer: Cyclic Prefix-OFDM used across the entire range of frequencies with scalable numerology and adaptive beamforming in higher frequencies, OQAM-FBMC supported at lower frequencies.
9. Adaptive Filtered OFDM with Regular Resource Grid: OFDM-based with flexible numerology and a regular resource grid, focus on massive MIMO and beamforming support, as well as asynchronous multiple access.
10. Harmonization aspects for D2D communications: focus on WFs and corresponding numerologies that are robust to some lack of synchronism (UF-OFDM and OQAM-FBMC).

To demonstrate the challenge of harmonizing features of different AIVs and integrating them into an overall AI, two of the above solutions (which take different approaches to AI design) and the relevant frequency/service mappings are described here in more detail. For full details of these and all other proposals the reader is referred to [MII16-D41].

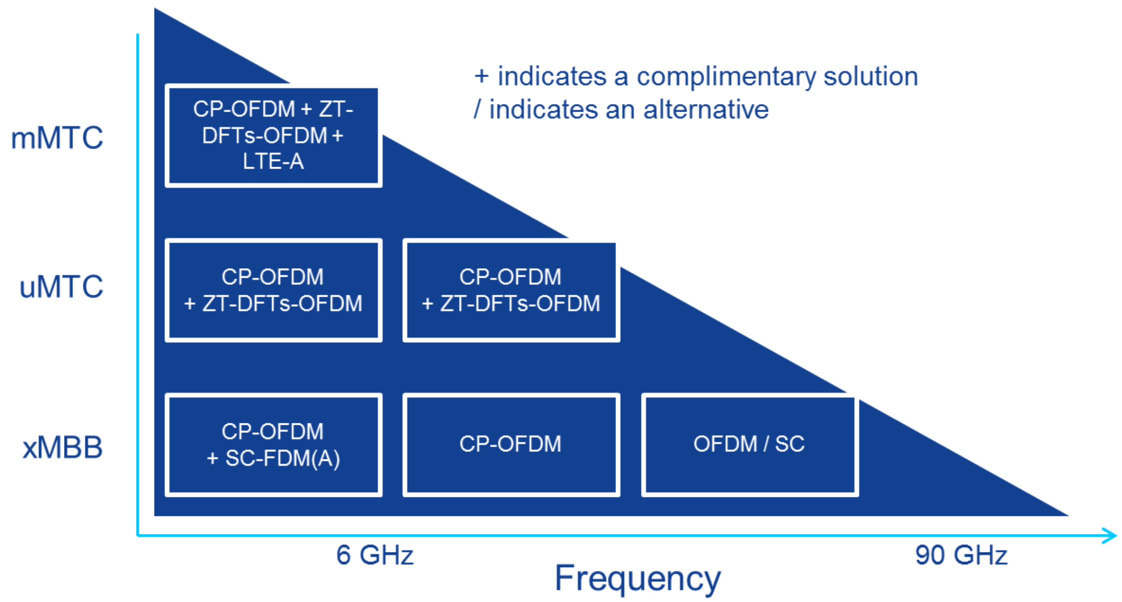


Figure 4-5. An OFDM-based AI framework with flexible numerology and frame structure. Cyclic prefix is denoted by "CP" in this figure.

In Figure 4-5, the frequency / service mapping for proposal 2 above is shown, as an example of a single WF family solution, with a high degree of similarity across PHY layers of component AIVs. This particular proposal utilizes Cyclic Prefix-OFDM(A) and its single carrier variant SC-FDM(A) (also known as DFT-s-OFDM(A)) with modifications to cater to multiple service types, as shown in the figure. As can further be seen, the main technology components are derived from OFDM and its enhancements, while also harmonizing the AIV characteristics such as sub-frame structure across different frequency ranges. As is well known from LTE-A evolution, Cyclic Prefix-OFDM based solutions can support a high degree of UP aggregation features, such as carrier aggregation. In this specific 5G AI proposal, these concepts are extended further. A single waveform family approach can be optimized for service types and use cases based on modifications such as a parametrized filter length and an adjustable zero tail, zero head length in the case of DFTs-OFDM. OFDM waveform parameters such as subcarrier spacing and guard bands in the time frequency grid can be chosen for optimized performance in a specific band. A frequency / service dependent numerology (with varying sub-carrier spacing) is proposed and described in more detail in [MII16-D41]. The solutions proposing the use of OFDM to address all 5G service types facilitate the usage of a common RM framework in a short time scale where for each transmit time interval (TTI), transport blocks can be scheduled for one service or another, or for uplink and downlink, depending on the traffic demands of each service.

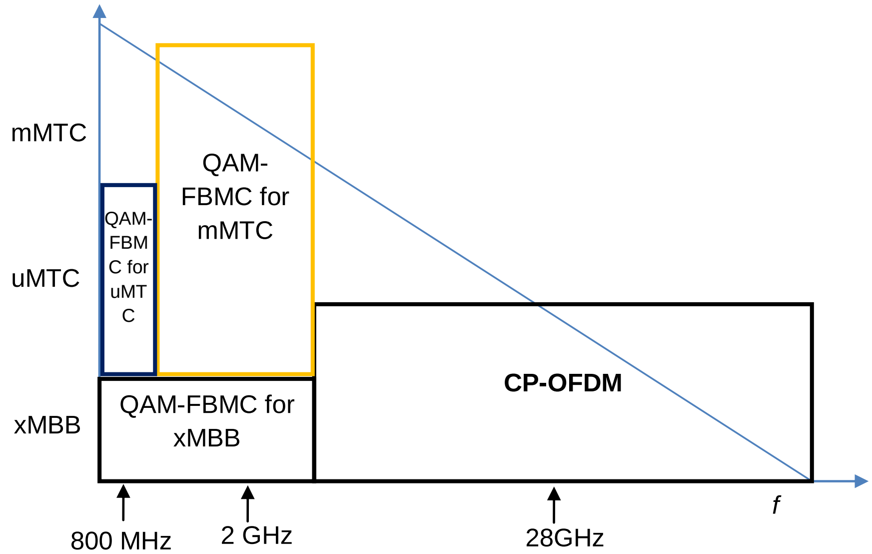


Figure 4-6. A multi-waveform AI framework with harmonization above the PHY. Cyclic prefix is denoted by "CP" in this figure.

In Figure 4-6, the frequency / service mapping for proposal 6 above is shown, as an example of a multiple WF family solution, with a lower degree of similarity across PHY layers of component AIVs. Two AIVs with different PHY features (in this particular case QAM-FBMC and Cyclic Prefix-OFDM) would need to be harmonized to support heterogeneous services that have quite different requirements. More specifically, when considering xMBB services with certain coverage requirements, e.g., in urban macro deployments, lower frequency bands are needed, and because of the more efficient usage of spectrum due to low out-of-band (OOB) emission, QAM-FBMC is favorable in such cases. Cyclic Prefix-OFDM is well-localized in time domain and thus suitable for delay-critical applications and TDD-based deployments. Even though OFDM has comparatively high OOB leakage (which can be tackled with enhanced OFDM solutions), this disadvantage can be easily compensated by the large available bandwidth in mmWave bands, e.g., 28 GHz. The frame structure and numerology for QAM-FBMC is however expected to be different because of the different symbol architecture and duration. Nevertheless, design parameters of Cyclic Prefix-OFDM can be highly flexible, and QAM-FBMC provides the additional degree of freedom of the filter design, which makes UP aggregation possible across the two AIVs. Additionally, and even though multiple waveforms are chosen, they could share the same modulation (in this case QAM) and coding schemes, e.g. low-density parity check (LDPC) codes. Moreover, a harmonized solution for reference signal placement, and baseband signal processing related to detection and equalization could be designed and is the subject of further study.

More details on these two examples and all other METIS-II AI proposals are available in [MII16-D41], where additionally an analysis of the commonalities of the proposals is provided, as summarized in Table 4-1 below.

Table 4-1. Commonality analysis of the METIS-II AI proposals.
Cyclic prefix is denoted by "CP" in this table.

	1. Harmonized L1-L3 with CP-OFDM	2. Harmonized CP-OFDM	3. UF-OFDM	4. Qualitative AIV design	5. P-OFDM	6. Harmonized QAM-FBMC & OFDM	7. OFDM with flex. numerology	8. Harmonized OQAM/FBMC & OFDM	9. Adaptive Filtered OFDM	10. Harmonization for D2D comm.
OFDM framework (O) or Multiple WFs (M)	O	O	O	M	O	M	O	M	O	M
Time localized symbols for xMBB	X	X	X		X	(x)	X	X	X	(x)
Time localized symbols for uMTC	X	X	X	X			X		(x)	(x)
Time localized symbols for mMTC	X	X	X	X	(x)		X		(x)	X
Scalable numerology over frequency bands	X	X	X	X	X	X	X	X	X	X
Symmetric waveform for UL/DL for xMBB	X		(x)		X	(x)	(x)	X	(x)	
Symmetric waveform for UL/DL for uMTC	X		(x)	(x)	X	X	X	X	(x)	
Symmetric waveform for UL/DL for mMTC	X		(x)	(x)	X	X	X	X	(x)	
Asynchronous communication for xMBB			X		(x)	(x)			X	(x)
Asynchronous communication for uMTC		X	X	X	X	X		X	X	(x)
Asynchronous communication for mMTC			X	X	X	(x)	X		X	X

The attributes being compared in the table are related to whether a solution is entirely based on the OFDM WF family (in other words, the OFDM framework with variations tailored to meet different 5G service requirements and bands), or whether it relies on multiple WFs (such as solutions that incorporate both OFDM-based and FBMC-based AIVs). It further captures whether the proposals are based on some key 5G AI design paradigms such as a scalable numerology, time localization, UL / DL WF symmetry, asynchronous communication, and how the support for these different paradigms varies for the three main 5G service types, for each of the proposals under consideration. An X in the table highlights that a feature is supported by the proposal, while an (x) means that this feature is covered only partially, as it applies as an option for selected use cases or a subset of the frequency bands only. For more details, please see [MII16-D41], where, following this analysis, an initial assessment was performed of the AI proposals (grouped around subjacent waveform technologies) using the harmonization KPIs conceived within METIS-II, where various forms of AI aggregation were identified and their features discussed.

The examination of the extent of harmonization inherent to various AI proposals also provides an initial overview of different possible UP aggregation approaches and various trade-offs each of these entails. Aggregation on a certain protocol stack layer means that on and above that layer there is only one single logical protocol stack instance, and hence the higher layers are agnostic with regard to the existence of multiple protocol stack instances at the lower layers. These aspects are covered in more detail in the context of the overall control and user plane architecture in Section 5.4.

While the above 5G AI proposals are not constrained to be backwards compatible with LTE-A, some benefits exist in harmonizing at least some 5G AI aspects with the LTE-A design, as examined in detail in the course of the work.

5 Overall System Architecture

The overall system architecture, from a logical perspective, is typically standardized in order to enable interoperability among multiple network manufacturers. The most fundamental part, which is also the focus of this chapter, is the mobile network architecture, which comprises both CN and RAN domains, the definition of network functions (NF, or network elements), standard interfaces and protocols running over these interfaces. More precisely, the chapter covers CN considerations, CN / RAN functional split, network interfaces and the protocol architecture for the AIs. The implementation or deployment of that logical architecture on a physical architecture comprises aspects such as backhaul, fronthaul, constraints in terms of hardware and software platforms, placement of the functions in the mobile network sites (access sites, aggregation sites, etc.), usage of cloud environments, centralization and distribution, etc. In addition to these two aspects, management and orchestration has gained a lot of attention in the past few years in the context of the 5G architecture. This has happened due to the expectations that at least a subset of the 5G NFs (e.g. CN) would be based on cloud platforms.

Regarding the deployment scenarios, it should be noted that both tight interworking between LTE-A and novel AIVs and new AIV standalone operation are considered in METIS-II, as shown in Figure 5-2. In particular solutions where LTE-A or novel AIVs could serve as anchor, providing coverage for the rest of the infrastructure, are being investigated, together with the standalone operation of novel AIVs, where these will provide full functionality without requiring or considering the interaction with LTE-A evolution.

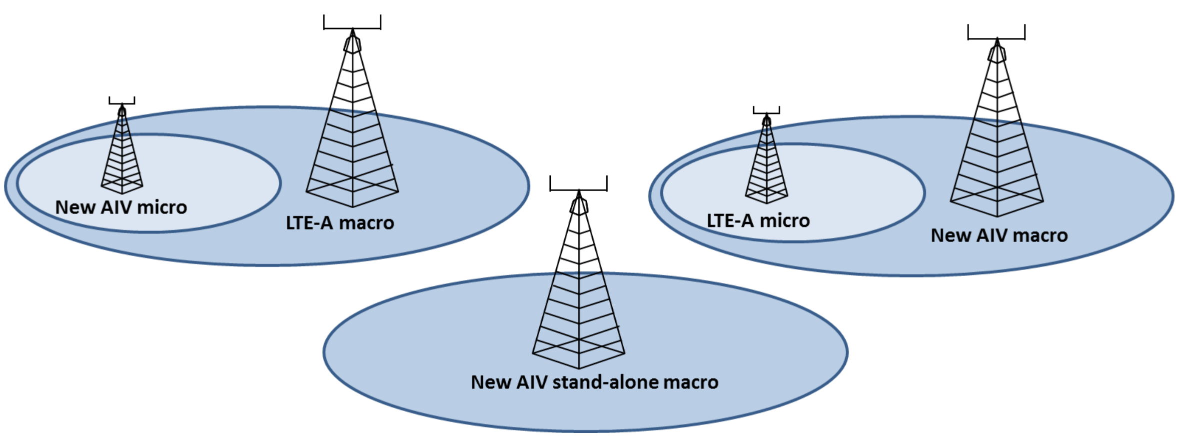


Figure 5-1. Example deployments of LTE-A evolution and a novel AIV.

5.1 Core Network

A detailed CN design is not in the main scope of METIS-II. However, in order to progress the overall RAN design, it is necessary to make some assumptions on CN functions and architecture, the CN / RAN split and related interfaces.

When it comes to the architecture support for novel AIVs and LTE-A evolution, METIS-II envisions an integration that goes beyond the existing interworking between access technologies. In existing systems, interworking of different radio technologies such as UTRAN and LTE relies on inter-node interfaces. This architecture basically allows internet protocol (IP) continuity, coverage continuity and load balancing only via hard handovers (i.e. always involving CN signaling) and semi-independent RM for the different access technologies [3GPP15-36300], which is clearly suboptimal.

An initial step towards an enhanced interworking is the assumption that both LTE-A evolution and novel AIVs could benefit from common 5G CN functions and the definition of a common CN / RAN interface). METIS-II foresees that these common functions will have the following characteristics:

- CN functions are expected to be deployed as virtual network functions (VNFs) in the 5G timeframe thus running in virtual machines over standard servers, potentially on cloud computing infrastructures – data centres. To which extent this Network Function Virtualization (NFV) could be extended to RAN functions is one of the key questions to be answered by METIS-II;
- The design of these CN functions will at some extent explore Software Defined Networking (SDN) principles such as UP/CP split, partially fulfilling the envisioned SDN/NFV native architecture;
- These VNFs (CP and UP) can be flexibly deployed in different sites in the operator's network depending on the requirements with regards to latency, available transport, processing and storage capacity, etc.;
- Different Services or Network Slices can utilize different CN and Service Layer VNFs which can be deployed at different network sites.

Some of these characteristics related to the cloud-based implementations of CN functions will require new management paradigms relying on some sort of orchestration framework to setup and manage these virtual machines. The METIS-II considerations about the management framework are provided in Section 5.7.

5.2 CN / RAN split

Another important assumption taken in METIS-II is the logical split between the RAN and CN (and Service Layer) functions. This is seen as beneficial for the following reasons:

- Allow for an independent evolution of RAN and CN functionality in order to speed up introduction of new technology;
- Enables to make the CN functions or a subset of them independent of the access (e.g. common UP processing);
- Facilitates mobility since some CN functions (CP and / or UP) can be kept (anchored) when UEs move to another RAN node;
- Allows cross-layer optimizations in some deployments when the functions are co-deployed, i.e. located in the same physical entity;
- A logical separation also facilitates multi-vendor CN / RAN interoperability.

The METIS-II baseline for the functional split between the RAN and CN is the same as in the EPS. However, alternative CN / RAN splits and/or cross layer optimizations are currently being studied. To give one example, METIS-II is currently investigating the design of RAN-based paging solutions to address densified deployments and a connected inactive state optimized for inactivity periods between small packet transmissions, along device mobility without involving CN / RAN signaling. Both approaches hence imply a shift of functionality from CN to RAN. More details can be found in Sections 6.3.3 and 6.3.2, respectively. The highlighted assumptions are also endorsed in the recently approved SA2 study item about the 5G architecture entitled “S2-153651 Study on Architecture for Next Generation System” [3GPP15-153651].

It should be emphasized that the considered CN / RAN split is a **logical** split and is mainly a standardization practicality to enable a multi-vendor ecosystem where one vendor can provide the CN, and another vendor the RAN, with both sub-systems operating well with each other. This does not forbid at all any sort of smart implementation on the network side if a single manufacturer builds both CN and RAN domains in the same network. For instance, a manufacturer may design both CN and RAN functions as a single solution, collapse these, split these into CP and UP, co-locate these etc. for joint optimization, as illustrated in Figure 5-2.

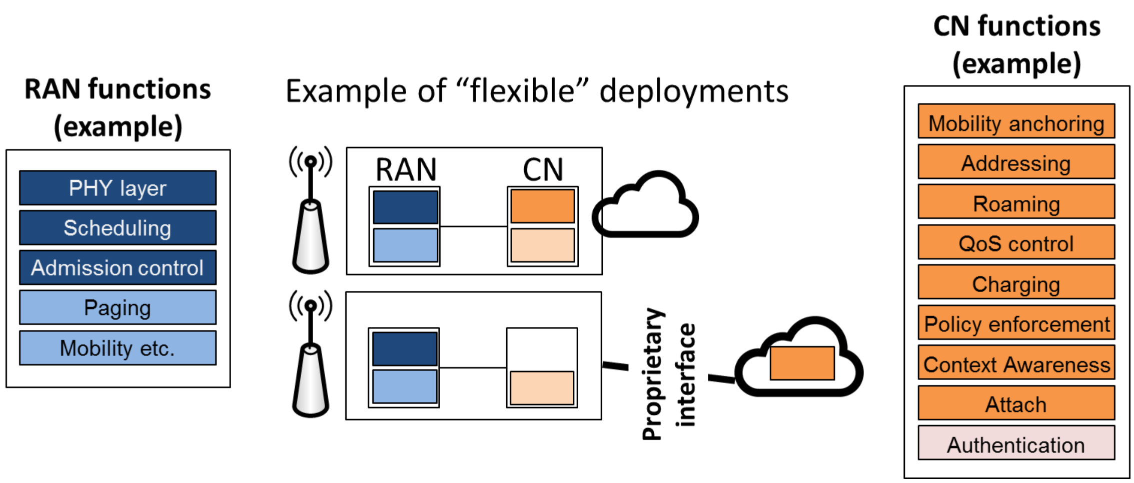


Figure 5-2. Possible flexible deployment and co-optimization of CN and RAN functions.

5.3 Network Interfaces

METIS-II proposes an integration of novel AIVs and the evolution of LTE that goes beyond the existing interworking between access technologies, fulfilling the vision of what NGMN calls a "5G RAT family". That should enable a high performing multi-RAT mobility between the new AIVs and LTE-A evolution and at least UP aggregation inspired in the Dual Connectivity solution. Considering these requirements and the previous assumption that different accesses could benefit from common 5G CN functions, it seems reasonable to assume a common CN / RAN interface for the new AIVs and the LTE-A evolution. METIS-II acknowledges that this CN / RAN interface, denoted herein S1*, will have a substantial amount of novel features designed to address the new future demands for the 5G architecture. For instance, it is envisioned that this will require the support for:

- E2E Network Slicing (where each slice may have its own set of CN functions);
- New 5G services with diverging requirements (e.g. with service-optimized CN functions);
- Enhanced multi-RAT integration with common CN functions where some could be designed to be independent of the access;
- Potentially new UP / CP splits in the 5G CN (designed to follow an SDN / NFV-native architecture, see Section 5.7);
- New connected state, optimized for battery savings but enabling fast transition to active.

A common CN/RAN interface has many benefits, such as:

- It makes it possible to very quickly establish dual connectivity for a UE first connected to a single RAT since there is no need to perform any extra CN/RAN signaling or non-access stratum (NAS) signaling when adding the second RAT
- It makes it possible to have a common evolution of LTE and novel AIVs where new CN features will benefit both RATs at the same time avoiding separate specification work.
- It simplifies the UE implementation since a single NAS layer is needed for both LTE and novel AIVs, hence avoiding a dual protocol stack at the UE.
- It simplifies the RAN / CN interaction since a single CP connection is used. This gives clear advantages when handling:
    - **Mobility**: a single handover procedure will be able to move the connections a UE has with each active radio accesses;
    - **State transitions**: Only a single EPS Connection Management (ECM) state needs to be kept. UE, RAN and CN behavior due to such single state are greatly simplified and the risk of losing state synchronization is reduced.
- Other signaling: a single CP connection avoids possible race conditions and error cases occurring if signaling is run over two independent connections.

To facilitate this envisioned integration of LTE-A with novel 5G AIVs, METIS-II assumes that this new interface will have the S1 interface [3GPP15-36300] as baseline. A main challenge needing further investigation is how to align the CN / RAN signaling evolution of LTE-A and novel AIVs.

Within the RAN, METIS-II initially envisions a new logical inter-node RAN interface, denoted herein X2*, designed to address possibly new features (see Chapter 6) such as:

- Intra-RAN mobility, also covering mobility between LTE-A evolution and novel AIVs;
- Multi-connectivity in non-collocated deployments of nodes from the same or different AIVs (e.g. novel cmWave and mmWave AIVs or LTE-A evolution);
- Support for RAN-based paging, where a tracking area may encompass multiple AIVs;
- Support for functions related to state transitions from Inactive to Connected such as context fetching;
- Support for novel interference management schemes;
- Support for RAN moderation to increase overall network energy efficiency.

Another aspect that needs to be taken into account is the fact that in addition to distributed deployments it should properly also support centralized ones, i.e., a centralized control of multiple synchronous functions (radio link control, RLC / medium access control, MAC / physical layer, PHY) by centralized asynchronous functions (packet data convergence protocol, PDPC and / or radio resource control, RRC) that could be possibly implemented in a centralized cloud. To facilitate the envisioned multi-RAT integration with LTE-A evolution, METIS-II assumes that this new interface will have X2 [3GPP15-36300] as its baseline. Details about the new interfaces are still being researched in METIS-II. The current working assumption on the logical 5G RAN architecture is captured in Figure 5-3.

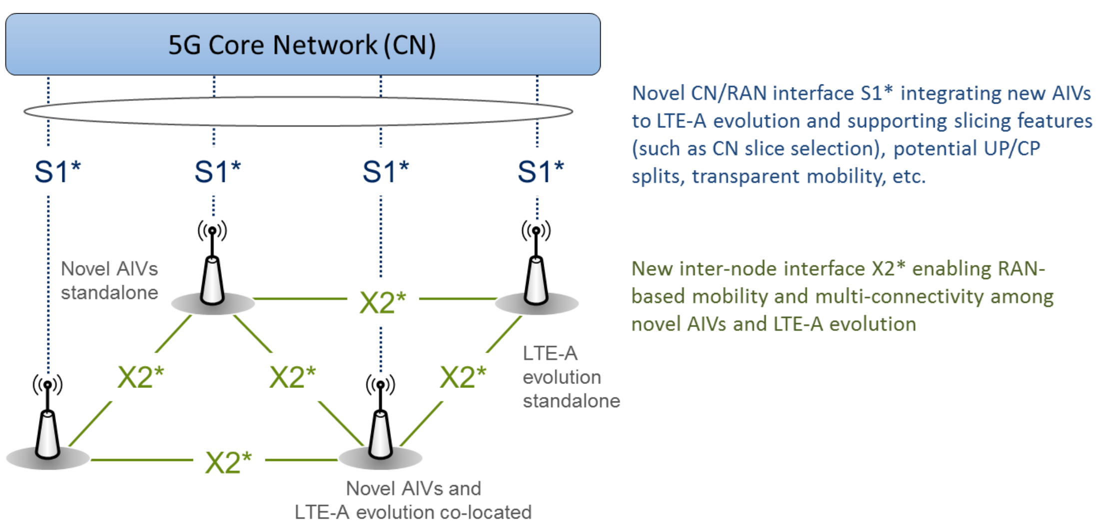


Figure 5-3. Current working assumption for the logical RAN architecture.

Considerations on RAN internal interfaces

During the early standardization phase of LTE, there was a clear requirement from operators that LTE should have a flat all-IP architecture. The background for this was that operators were not satisfied with the Universal Mobile Telecommunications System (UMTS) architecture splitting RAN level functionality between the RNC and NodeB. Issues with the radio network controller (RNC) included things such as split of ownership of resources between NodeB and RNC and RLC termination point in the RNC, putting strict requirement on flow control between the NodeB and RNC. Furthermore, the UMTS architecture required substantial changes during the standardization of High Speed Downlink Packet Access (HSDPA), where some functionality originally introduced in the RNC was moved to the NodeB (such as MAC scheduling).

The outcome of the research done at that time was an LTE architecture comprised of a single logical RAN node (eNB) which terminates the S1 and X2 interfaces. Nevertheless it has been shown that the LTE architecture does allow for a high level of architecture flexibility, e.g. towards centralized baseband pooling, e.g., facilitating coordinated multipoint (CoMP) solutions across several transmission points. It is also possible in the implementation for example to further centralize and virtualize higher layers of the radio access protocol stack if that is found to be beneficial.

It is also worth noting that a lot of new features (e.g., Multimedia Broadcast Multicast System (MBMS), home eNB, self organizing networks (SON) enhancements, IP multimedia sub-system voice over LTE (IMS VoLTE), Improved Access Control, Mobility enhancement, carrier aggregation, dual connectivity, machine-type communications (MTC) enhancement, Wireless Local Area Network (WLAN) integration, Relays, Proximity based services, D2D communications, Positioning, MIMO enhancement, user equipment (UE) specific demodulation reference signal (DMRS), CoMP, enhanced inter-cell interference coordination (eICIC), Network sharing, etc.) have been added to LTE since Rel-8 without requiring any fundamental changes to the RAN internal architecture as specified in 3GPP. This flat RAN architecture has therefore accelerated the standardization of the above-mentioned functionality. More importantly, it also simplified the implementation and in particular the inter-operability-testing.

If Evolved-Universal Terrestrial RAN (E-UTRAN) would have been split into several RAN internal nodes (as formerly done in UTRAN), the interfaces and the functional split between these nodes would have also been impacted leading to the following problems:

- Delayed standardization of new radio features since also internal interfaces would have needed to be standardized;
- Delayed and more complex implementation and testing of the new features;
- A risk that early decisions of the functional split between the internal nodes would have been sub-optimal for later features introduced in LTE which either would have meant that sub-optimal solutions would have had to be adopted, or a major redesign would have been needed;

- Different vendor’s implementations could have made it difficult to agree on the preferred split, potentially leading to sub-optimal solutions being adopted in the standard. Such suboptimal solutions could have resulted in lack of inter vendor interoperability and higher integration costs.

The UMTS architecture on the other hand has attempted to specify a protocol split by standardising the Iub interface between NodeB and RNC. This has proven to be rather inefficient, some of the issues being as follows:

- One of the functions of the Iub interface is to allow Radio RM (RRM) from the RNC to the NodeB. RRM is made of proprietary algorithms involving processes shared by the RNC and NodeB. It is extremely difficult to have an efficiently working Iub interface in a multi-vendor RNC-NB deployment because of the tailored vendor specific processes each node would support, and that cannot be supported by a standardized interface;
- The processes run over the Iub can be very delay-sensitive, examples can be scheduling coordination, UL/DL power control etc. In situations where the RNC-NB connection is not sufficiently performing it is very difficult to make the Iub operate in an efficient way. In such scenarios a different RAN architecture split would have been more suitable.

The issues above related to the UMTS technologies revealed to be difficult to solve, which is why the work on a flat LTE architecture was triggered and why this choice has been made. In the 5G era, it is important to have this history of architecture approaches in mind when deciding on novel RAN-internal interfaces.

5.4 Overall Control / User Plane Architecture

This section is split into two parts: Section 5.4.1 investigates to which extent protocol functions in the new AIVs would need to differ from the current ones in LTE-A (taking Release-13 as a reference) to meet the requirements in 5G. It also elaborates on the question to which extent one would likely have different specifications for LTE-A evolution and the novel AIVs. Sections 5.4.2 and then venture into the overall control / user plane architecture as such, for the case of the LTE-A evolution integration with novel AIVs, or the integration of novel AIVs among each other, respectively.

Before starting, it may be valuable to say a few words about the definition of “user plane” and “control plane” as such: The functions or protocols designed to carry end-user data (such as IP packets) are typically called UP functions, while functions related to control are called CP functions. In this respect, the RRC, responsible for functions such as mobility control, connection control and system information transmission / acquisition, is a clear CP functionality. The PDCP, however, is predominantly used for compression and decompression of IP data flows, transfer of data, security, and maintenance of PDCP sequence numbers etc., hence aspects related to UP processing, while at the same time also being used to convey RRC-related information from and to the lower layers. The same applies to the MAC and physical PHY, which are relevant for both CP and UP. For this reason, the CP protocol stack for the RAN is typically drawn from PHY up to RRC, and the UP protocol stack from PHY up to PDCP.

5.4.1 Possible Changes in Protocol Functions for the new AIVs w.r.t. LTE-A, and the Usage of Common Specifications

A key question in METIS-II is to which extent protocol functions of the new AIVs may have to be substantially modified to meet the 5G requirements. In this subsection, we will hence explore the different protocol stack layers, list their current functions as in LTE-A, and elaborate on any potential changes in 5G. From this, we can also derive the notion of whether LTE-A and new AIVs should use common specifications in the future.

Radio Resource Control (RRC)

In LTE-A, the RRC functions are responsible for the broadcast of system information for NAS and access stratum (AS), paging, connection handling, allocation of temporary identifiers, configuration of lower layer protocols, quality of service management functions, security handling at the access network, mobility management, and measurement reporting and configuration, etc.

New enhancements are expected to be part of the LTE-A evolution, such as the lightweight connection Work Item based on the Suspend/Resume procedure and possibly some paging enhancements [3GPP15-23720]. However it is not clear whether that will lead to RRC changes or not for the LTE-A evolution.

For the new AIVs, it is expected that the role of the RRC protocol will in general stay the same and most of the previously mentioned functions would still be necessary. Having the same role, however, does not necessarily mean the same exact design. One potential change compared to LTE-A RRC would be the need to **support beam-based measurements and reporting mechanisms** (other levels might exist as in LTE-A e.g. based on PHY layer and channel state information-reference signal, CSI-RS), see Section 6.1.2. In addition, the design of **new ways to distribute and encode system information** is also being considered, as elaborated in more detail in Section 6.1.3. Another considered change is the introduction of a **new RRC state** (see Section 6.3.1), but since the lightweight connection WI [3GPP15-23720] is still ongoing, one cannot really say how disruptive this will be compared to LTE-A evolution. Another change is the support for tight interworking between LTE and the new AIVs, including mobility of active/inactive UEs and UP aggregation (more details in Section 5.4.2). Since LTE-A evolution and the new AIVs should be possibly supported as anchors, an RRC instance needs to allow the establishment of a secondary link of another AIV and vice versa for UEs having one of the new AIVs as anchor.

All these potential design changes seem to justify the definition of a **new RRC specification** for the new AIVs, although they might use the LTE-A RRC specifications as its baseline. However, due to the tight interworking between LTE and the new AIVs, it is very likely that at least some level of interaction between the specifications will be necessary, which is to be further researched.

Packet Data Convergence Protocol (PDCP)

In LTE-A, PDCP is responsible for compression and decompression, transfer of UP and CP data, security (i.e. encryption), maintenance of sequence numbers etc. An overview on all PDCP functions and possible changes in 5G is provided in Table 5-1.

Table 5-1. PDCP functionalities in LTE-A and possible changes in 5G.

Functionality in LTE-A (Release 13)	Considerations for novel AIVs in 5G
Compression and decompression of transmission control protocol /user datagram protocol/IP (TCP/UDP/IP) headers, which is essential for small payloads where the relative header overhead is large.	For novel AIVs, compression and decompression may be **tailored to different services**. For instance, header compression may be omitted if the payload is fairly large and/or latency is crucial (e.g. some xMBB or uMTC applications), while being more pronounced in the context of mMTC.
Security functions, e.g. ciphering and deciphering of UP and CP data. Integrity protection and integrity verification of CP data. For relay nodes, integrity protection and integrity verification of UP data.	No changes foreseen. PDCP is still seen as the most suitable layer for ciphering / deciphering, as in 5G deployments the lower layers may more often be placed in user-deployed entities which may be compromised.
Maintenance of PDCP sequence numbers (SNs), duplicate detection/elimination and discarding, and timer-based discard.	No changes foreseen.
Routing and reordering of PDCP protocol data units (PDUs) in the case of split bearers (RLC acknowledged mode, AM).	No changes foreseen. This functionality is seen as particularly important for the widespread usage of multi-connectivity in 5G.
Data-recovery procedure for split bearers in DC (for RLC AM), for instance needed when part of the data transmitted over one radio leg is lost due to bad radio conditions.	No changes foreseen, though in 5G the data-recovery procedure will need to be defined for both multi-connectivity among LTE-A evolution and novel 5G radio, as well as among multiple novel AIVs.
Retransmission of PDCP service data units (SDUs) at handover: The handover case is very similar to the use case for the data-recovery procedure.	No changes foreseen.

As can be seen from Table 5-1, there are no major changes expected to be introduced for PDCP, except a better tailoring to different services types, and the usage of the data-recovery procedure to also cover multi-connectivity among LTE-A evolution and novel 5G AIVs which might not require standard changes to PDCP. Since the PDCP layer will be an important layer for the aggregation of multiple AIVs, in particular the aggregation between LTE-A evolution and novel AIVs, as we will see later in Section 5.4, it is however important, that various forms of interworking between the LTE-A and PDCP specification are possible. For instance, if an LTE-A evolution radio and a novel 5G radio would be aggregated on PDCP level, this PDCP instance may follow the 5G specification and should hence also be able to utilize the services provided by the LTE-A lower layers. It may also be considered to use the LTE-A specification for the common PDCP instance, which should then also interoperate with the RLC, MAC and PHY layers of the novel 5G AIV, as discussed further in Section 5.4.2 and illustrated in Figure 5-4.

Radio Link Control (RLC)

For RLC layer, the main function is automatic repeat request (ARQ) and data segmentation/concatenation, based on which mode (acknowledged or unacknowledged mode) is configured. The following table summarizes the RLC functionalities defined in LTE in more detail, and elaborates on the potential changes considered by METIS-II partners for 5G.

Table 5-2 . RLC functionalities in LTE-A and possible changes in 5G.

RLC functionality in LTE-A	Considerations for novel AIVs in 5G
Transfer of upper layer PDUs	No change foreseen.
Error correction through ARQ (only for AM data transfer). By configuring AM RLC, ARQ is supported with an extra layer of retransmission reliability.	For novel 5G AIVs, the combination of ARQ and hybrid ARQ (HARQ) should be further studied. Since it may be possible to improve the reliability of MAC HARQ, the ARQ may in some use cases potentially be omitted.
Concatenation, segmentation and reassembly of RLC SDUs (only for unacknowledged mode, UM, and AM data transfer), for the purpose of generating RLC PDUs of appropriate size from the incoming RLC SDUs.	Since concatenation and segmentation require the knowledge on the MAC transport block sizes, this RLC functionality is tightly tied to the MAC and hence has to happen on synchronous time scale. A consideration is to move this functionality into the MAC, while keeping individual queues per RLC entity to avoid head-of-line blocking. This way, the remaining RLC functions would be asynchronous, and a function split between RLC and MAC would be a split between asynchronous and synchronous functions, see Section 5.5.2. It yet has to be clarified to which extent this would touch standardization, or be a matter of implementation.

Re-segmentation of RLC data PDUs (only for AM data transfer), in the case that these do not fit to the actual transport blocks.	In novel 5G AIVs, the usage of this function may be extended to new scenarios, for example, the usage of unlicensed spectrum, where the transmission may be blocked by channel acquisition. Then, the RLC PDU could be re-segmented to fit the next transmission.
Reordering of RLC data PDUs, duplicate detection and RLC SDU discard (only for UM and AM data transfer), **RLC re-establishment, and protocol error detection** (only for AM data transfer)	No changes foreseen.

In summary, the services and functions supported by RLC in LTE-A should be the baseline for the 5G design, but some specific changes are foreseen for novel 5G AIVs, such as the dynamic usage of ARQ depending on H-ARQ reliability and use case, a relocation of concatenation, segmentation and reassembly to the MAC, and the tailoring of re-segmentation to the usage of unlicensed spectrum. In this respect, it is clear that novel 5G AIVs will have a specification which is distinct from the LTE-A specification.

Medium Access Control (MAC)

The design of MAC in LTE-A has allowed keeping a low complexity with an efficient and fast handling by any type of devices of the transport block. This has been achieved by keeping a minimalistic approach to the packet handling functions that are mainly responsibility of RLC and concentrating in the functions that allows the optimal operation and utilization of the physical layer (this would of course change of part of the RLC functionality is moved to the MAC, see previous RLC description). The main services and functions of the MAC sublayer include:

- Mapping between logical channels and transport channels
- Multiplexing / demultiplexing of MAC SDUs belonging to one or different logical channels into / from transport blocks (TBs) delivered to / from the physical layer
- Priority handling between logical channels of one UE. The handling of different priorities and an efficient use of spectrum motivates to implement these functions in MAC.
- Initial Access using the Random Access Channel for requesting uplink resources.
- Scheduling information reporting. The reporting of the UE scheduling information is an efficient and fast function that allows the network to provide the UE with UL grants. In a fully scheduled system, this function will be still required for the same purpose.
- Error correction through HARQ. The benefits of HARQ retransmissions motivate the need to provide HARQ retransmissions in MAC for services that require high reliability.
- Priority handling between UEs by means of dynamic scheduling;
- Transport format selection;
- Padding.

These functions are foreseen to be also required by the MAC layer of new AIVs. Some of them might require adjustments in order to provide higher flexibility to address the requirements of the new 5G use cases. For example, in order to reach very high data rates with the deployment of wide carriers, a new set of transport formats are needed to be defined and possibly a new transport format selection procedure needs to be updated according to the newly defined control channels for new AIVs. Another example is the improvements to the UL granting signaling to enable a greater granularity and control of logical channels when multiplexing data. As mentioned during the description of the RLC layer, it is considered to move synchronous RLC functionalities such as concatenation and segmentation to the MAC layer, hence extending the list above. Furthermore, considering beam-centric design (see Section 6.1.2), MAC may also be involved in beam-related reporting, e.g., in case of dynamic traffic steering.

5.4.2 CP / UP Architecture for the Interworking of LTE-A evolution with novel AIVs

A tight interworking between LTE-A evolution and novel AIVs has been assumed in METIS-II from the very beginning of the project. These assumptions have been later adopted in 3GPP, where TR 38.913 describes that the 5G RAN should support high performing multi-AIV[4] mobility and UP aggregation as requirements for tight interworking [3GPP16-38913].

For active UEs, high performing inter-AIV mobility can be translated into high robustness against packet losses (lossless), handover (HO) and radio link failures (RLF); low interruption delays (seamless) and low signaling overhead in the radio interfaces (i.e. LTE-A and novel AIVs) and in the network side. METIS-II has concluded that a **high performing inter-AIV mobility between LTE-A evolution and novel AIVs should be realized on a RAN level**, i.e. without necessarily involving CN signaling.

A key technology in this respect will be multi-connectivity between LTE-A and novel AIVs. From a UP perspective, the most simple realization of such is to aggregate LTE-A and novel 5G AIVs on PDCP level, i.e. have a common PDCP instance for both. For this form of multi-connectivity, the lower protocol stack layers may have independent specifications, which was actually recommended in Section 5.4.1. Due to the aforementioned aspect that both LTE-A and novel AIVs could serve as an anchor layer, either technologies should be able to perform the PDCP flow split, and both the PDCP specification of the 5G AIV and that of LTE-A should be able to rely on services provided by the lower layers of the respective other technology. This is depicted in Figure 5-4.

[4] In 3GPP referred to as multi-RAT

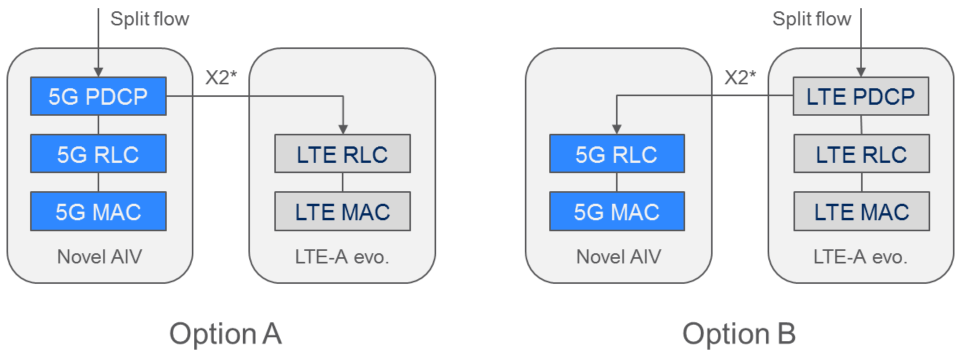


Figure 5-4. Multi-connectivity among LTE-A and a novel AIV.

Regarding the design of the RRC architecture to support this form of LTE-A evolution and novel AIV interworking, multiple alternatives are being considered, which are differing in whether there is a single or dual RRC connection, and whether there are one or RRC entities at the network side. When analyzing these options in detail, which we will do in the sequel, the following considerations are most relevant: the CN / RAN connectivity, the number of RRC connections / UE states, the number of RRC entities at the network, and the transport of RRC messages.

CN / RAN connectivity: In order to support this fast establishment from either LTE-A evolution or the new AIVs it is assumed a single RAN/CN connectivity, i.e. single NAS, for LTE-A evolution and the new AIVs. This solution seems to be reasonable considering the assumption of a common CN and a common CN/RAN interface for LTE-A evolution and the new AIVs. Such a solution makes it possible to have a common evolution for CN features of LTE-A and the new AIVs benefiting both AIVs at the same time avoiding separate specification work. In addition to that, a single NAS connection also simplifies the UE implementation, hence avoiding a dual protocol stack at the UE. When it comes to other procedures, it gives advantages when handling mobility and state transitions. In the case of mobility, a single handover procedure will be able to move the connections a UE has with each active AIV. In the case of state transitions, only one single CN state needs to be kept. UE, RAN and CN behavior due to such single state are greatly simplified and the risk of losing state synchronization is reduced. The main challenge is the need to align the CN/RAN signaling evolution of LTE and 5G.

Number of RRC connections / UE state(s): The project considers one alternative based on the Release-12 solution where a UE maintains a single RRC connection with a single MeNB and has a single RRC state and follows a single set of well define RRC procedures. In case of Dual Connectivity setup, the procedure could closely resemble the signaling in LTE-A. The solution also enables a lower complexity at the UE side since the second AIV do not require the establishment of a new state machine. In a second alternative, the UE maintains dual RRC connections with two eNBs and has two state machines running in parallel. One potential advantages of that approach is that the SeNB could configure its own resources directly by sending RRC reconfiguration to the UE, which in principle could reduce latency. However, some coordination should anyway occur between the different RRC entities from MeNB and SeNB anyway e.g., in order to sustain a single CN/RAN connection and coordinated state transitions between the two AIVs.

Number of RRC entities at the network side: it is very important to clarify that a single RRC connection and a single UE state does not necessarily implies a single RRC entity at the network side but rather implies that the UE only sees and communicates with a single entity. In other words, either for single or dual RRC connection, as described previously, there could be one or two RRC entities at the network side generating ASN.1 associated to each AIV. That may define the way one RRC specification understands the other.

Transport of RRC messages in Dual Connectivity: In the case of a single RRC connection, the MeNB generates only one final RRC message. In the case of two RRC entities at the network, there could be different ways to transport that final message associated to two different AIVs. One possible solution could be that information elements (IEs) associated to one AIV includes broadcasted or dedicated system information and security control information elements from the other, possibly generated in another eNB and coordinated over X2* (e.g., carrying radio resource control information elements). These messages can be carried within RRC containers that need to be specified.

Considering that the new AIVs operating in higher frequencies will rely on beamforming where fast SINR drops may occur due to link blockage and higher penetration loss, mobility robustness becomes more critical than in LTE-A so that multiple routing algorithms for RRC messages become more attractive. Among them, METIS-II considers RRC diversity both for UL and DL (or hybrids where only one direction benefits from diversity), fast switching, etc. Figure 5-5 shows different transport configurations for the two alternatives, single RRC and dual RRC.

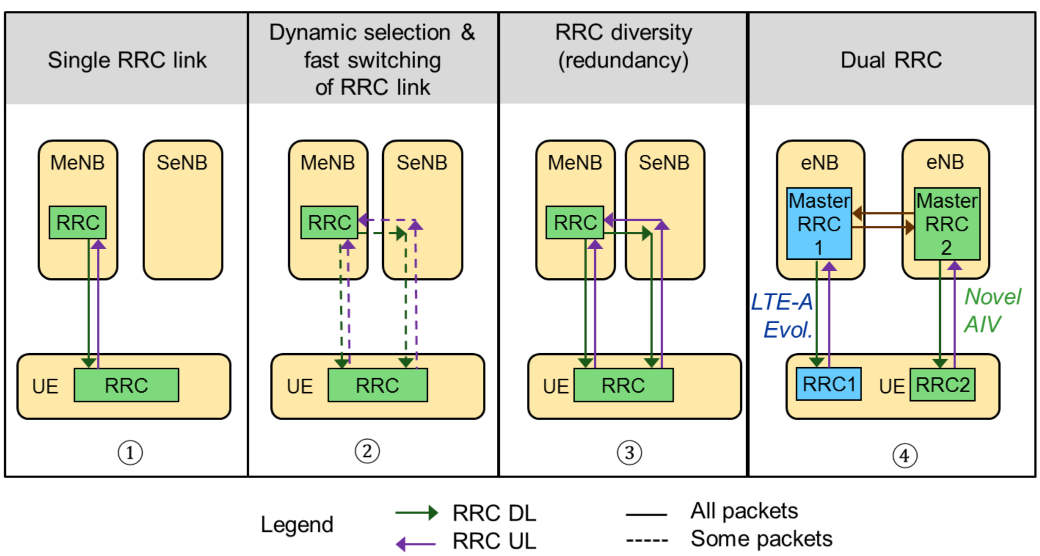


Figure 5-5. Different CP configurations with single RRC and dual RRC.

5.4.3 CP / UP Architecture for the Interworking of novel AIVs

The CP and UP architecture proposed for the interworking of novel AIVs is rather similar to that considered for the interworking of LTE-A evolution and novel AIVs, but is further facilitated by the fact that all novel AIVs can be natively designed to have a large extent of protocol harmonization and allow tight interworking. This will be visible through the following points:

- Due to the expected harmonization of the MAC and higher layer functionality across different AIVs, it may be possible – for instance in the context of co-deployment or the usage of cloud-RAN – to aggregate the UPs of different novel AIVs on MAC level, such as in LTE Rel. 10 carrier aggregation;
- For the CP architecture, it is in general expected to have one RRC connection for any UE served in multi-connectivity across novel AIVs. As for the interworking between LTE-A evolution and novel 5G AIVs, it is most likely that there will be multiple RRC instances at the network side, though ultimately only one RRC instance generates the RRC messages to be sent to the UE.

It has to be noted that beyond the architecture support for AIV interworking, METIS-II is developing various functional concepts to allow leveraging the nature of different AIVs, for instance related to traffic steering and RM in Section 6.2 and initial access and mobility management in Section 6.3.

5.5 Physical Architecture and Function Deployment

5.5.1 Deployment Scenarios Considered

METIS-II considers that the 5G RAN has to be designed such that it supports a range of different physical architectures. In particular, the RAN design should be able to leverage centralization in scenarios where for instance dark fiber is available to the edge, but also work well in scenarios of distributed base stations with possibly limited backhaul infrastructure. To evaluate 5G RAN concepts developed in METIS-II w.r.t. their suitability for different physical architectures, METIS-II has chosen four representative physical architecture deployment scenarios depicted in Figure 5-6. These scenarios have been chosen and defined such that

- they reflect in a simplified form the corner cases of deployment that can be expected in a 5G time frame (e.g. deployment of novel 5G radio sites in addition to existing LTE-A sites, co-deployment of LTE-A and new AIVs, wide-spread usage of self-backhauling, stand-alone sites and centralization of baseband processing etc.), and
- they do not imply any particular logical architecture / protocol stack architecture, as this is the subject of research in METIS-II itself.

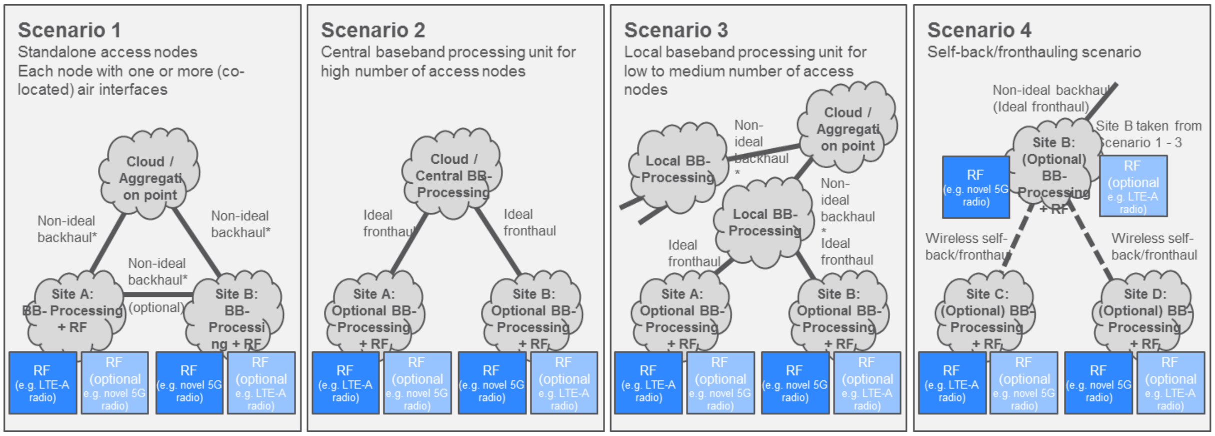


Figure 5-6. Physical architecture / deployment scenarios considered.

The scenarios are detailed in the following:

- **Scenario 1** depicts a case where the same or different AIVs are served from two (or more) spatially separated sites, and where the connection between the sites and from the sites to the next aggregation point are prone to non-ideal backhaul, which could for instance be modeled as in [3GPP14-36932]. For simplicity, it is assumed that CN functionality also resides at the aggregation point, hence there is no further latency involved beyond the aggregation point. As an option, multiple AIVs (e.g. LTE-A and a novel AIV) could be served in a co-located way.
- **Scenario 2** depicts the case where centralized processing is applied to all cells, and different spatially separated access nodes are served with ideal or non-ideal fronthaul, depending on the RAN functional split that is deployed.
- **Scenario 3** is similar to Scenario 2, but considers multiple clusters of centralized processing, each serving multiple radio sites that are assumed to be connected via ideal fronthaul. The connections between the centralized processing clusters, however, are assumed to be prone to non-ideal backhaul.
- **Scenario 4** builds upon Scenarios 1-3, but now depicts a case where two (or more) additional access nodes establish wireless backhaul links to sites with wired backhaul.

The intent is that **all concepts and frameworks developed in METIS-II, and the overall RAN design, must be capable to support and leverage from all four physical architecture scenarios**.

5.5.2 Possible Function Splits and related Intra-RAN Interfaces

In order to analyze how the logical METIS-II architecture described in Section 5.4 can be best mapped to the physical architecture, i.e. the deployment scenarios listed in the previous section, it is important to consider how NFs can be split over different physical entities, and which intra-RAN interfaces between the physical entities would correspondingly be needed.

The key rationale behind any choice of function split is to obtain the largest possible extent of centralization that a specific deployment architecture supports. A large extent of centralization of functionalities allows to exploit gains related to, e.g., centralized joint transmission, centralized scheduling, centralized flow control etc., but the price are increased fronthaul data rate requirements and increasingly stringent latency requirements, as we will discuss later. Figure 5-7 shows key split options, based on previous work in [NGM15-CR, MET15-D64, RBM+15, iJOIN-D22, iJOIN-D23], but also extended by latest considerations in METIS-II and CHARISMA [CHAR16-D11].

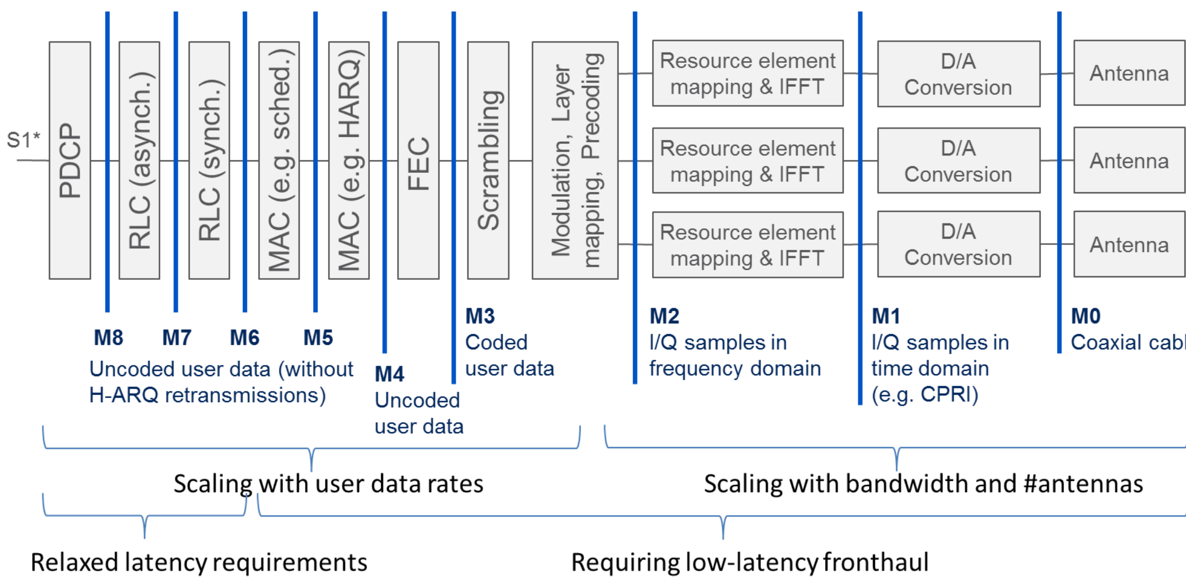


Figure 5-7. Function split options possibly implemented within the UP protocol stack.

One key aspect in this respect is the **data rate** required on the resulting fronthaul interfaces, for instance between a Remote Radio Unit (RRU) at an antenna site and a Baseband Unit hosting the full radio protocol stack or upper parts of it in a decentralized or centralized way (cloud/centralized-RAN, C- RAN). This data rate is typically predominantly driven by the amount of UP data exchanged. The basic data rate scaling behavior for the possible function splits is the following:

- M1: Here, the interface carries the digital baseband data in time domain for each antenna port. The required data rate hence scales with **system bandwidth**, number of

quantization bits per in-phase /quadrature (I/Q) sample, and **number of antennas**. Especially in 5G, where large system bandwidths and a large number of antenna ports may play a role, such interface may be prohibitive in required data rates.

- M2: Here, also digital base band data is conveyed per antenna port, but in frequency domain, which has the benefit that data only needs to be transported for sub-carriers which are actually scheduled and used for transmission (and not, e.g. guard carriers). Further, a lower number of quantization bits per I/Q samples can be used. The scaling behavior is otherwise the same as for M1.
- M3: Here, the information conveyed is the coded user data before scrambling, modulation and layer mapping/precoding. Hence, the data volume here scales solely with the **user data rates and the selected forward error correction (FEC) code strength**, and not strictly with the system bandwidth, number of antenna ports etc.
- M4: This is essentially the same as M3, except that the uncoded user data before coding is conveyed over the fronthaul interface. The scaling behaviour as such is the same as before.
- M5-M8: Here, the information conveyed on the intra-RAN interface is the uncoded user data without HARQ retransmissions, again scaling with the amount of user data. In terms of data rate requirements, M5-M8 are equivalent, but the interface options differ strongly in latency requirements, as we will discuss later.

An initial analysis of the interface bandwidth requirements for the different split options is shown in Figure 5-8, here based on the assumption of an LTE numerology of a 20 MHz system with 2x2 MIMO. Details about the models can be found in Appendix A.1.

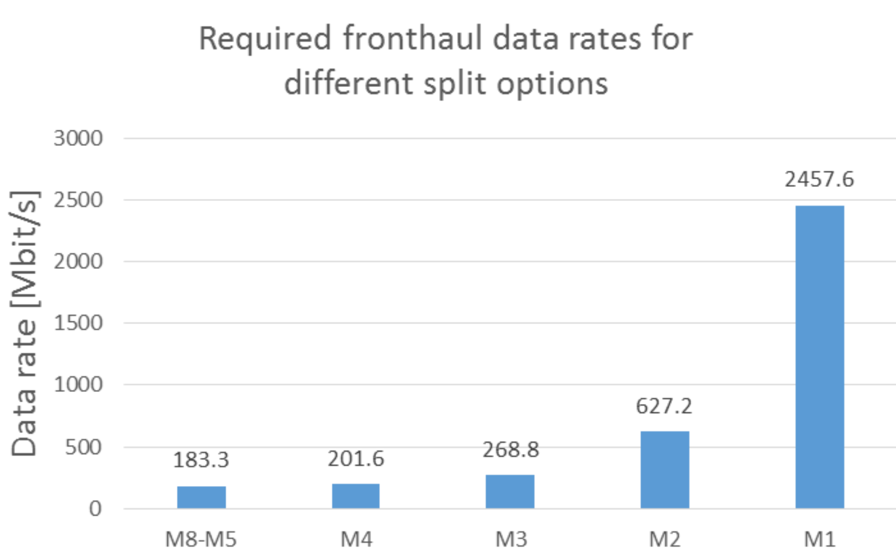


Figure 5-8. Fronthaul data rate needs for different UP split options (assuming 20 MHz system bandwidth and 2x2 MIMO).

Results show that split option M1 requires significantly higher data rates compared to the other split options. Compared with M2, M1 requires approximately four times higher data rates. This can be explained by the fact that M1 has to carry the data after/before the inverse fast Fourier transform/fast Fourier transform (IFFT / FFT). This means that I/Q samples for all subcarriers after transformation in the time domain have to be transmitted, also for the subcarriers carrying no data. The FFT size (i.e., the overall number of subcarriers) is 2048. Considering also 144 samples for the cyclic prefix results in a total number of 2192 samples per OFDM symbol, while the number of subcarriers carrying data only corresponds to 1200. Another difference comes from the number of quantization bits for each signal component. For M1 a value of 15 is used and for M2 a value of 7.

In addition to data rate requirements, also **latency** aspects are a critical issue for the selection of suitable splits, for instance limiting the implementation of certain functionalities (e.g. CoMP processing) in the case of some deployment scenarios. In this respect, preliminary analyses [MET15-D64] have concluded that in particular in the context of latency-prone backhaul / fronthaul, as in deployment scenario 1, time-*synchronous* functions (in LTE these are PHY, MAC and RLC functions such as scheduling, link adaptation, power control, interference coordination etc.), should ideally be placed close to the radio units. Many of these functions are also difficult to virtualize, as they often depend on hardware acceleration. Functions which are time-*asynchronous* to the radio interface (in LTE these are PDCP and RRC functions related to measurement control and reporting, handover preparation and execution, dual connectivity, random access, RRC state transition etc.), however, could be implemented as VNFs and possibly centralized, as they can typically cope with larger latency (e.g. tens of milliseconds in LTE-A).

In this context, a key consideration is to design 5G RAN functions to avoid strict timing relations between the protocol layers, and have a clearer split between time-synchronous and time-asynchronous functions. One specific design consideration in 5G, as already stated in Section 5.4.1, is for instance to move the time-synchronous functionalities of segmentation and concatenation from the RLC layer to the MAC layer, such that the RLC only contains time-asynchronous functionalities. This way, the split between RLC and MAC could be a suitable point for an intra-RAN interface, in particular in the context of a latency-prone fronthaul / backhaul infrastructure. This would correspond to split option M7 in Figure 5-7. If a reliable low-latency fronthaul is available, a viable option would be split option M5 – this allows to centralize upper MAC functions such as scheduling, but still substantially relaxes the fronthaul bandwidth requirements as compared to the state of the art, i.e. a Common Public Radio Interface (CPRI) interface (split option M1).

5.5.3 Specific Mapping Considerations of Logical to Physical Architecture in METIS-II

As an abstract function split model applicable to all physical deployment scenarios, METIS-II is considering the model depicted in Figure 5-9. As one can see, the consideration is to generally split the NFs in the RAN into a lower, middle and upper layer. This split may be done differently for UP and CP. Let us now see how the splits may be realized for the specific physical deployments considered in METIS-II in Figure 5-6.

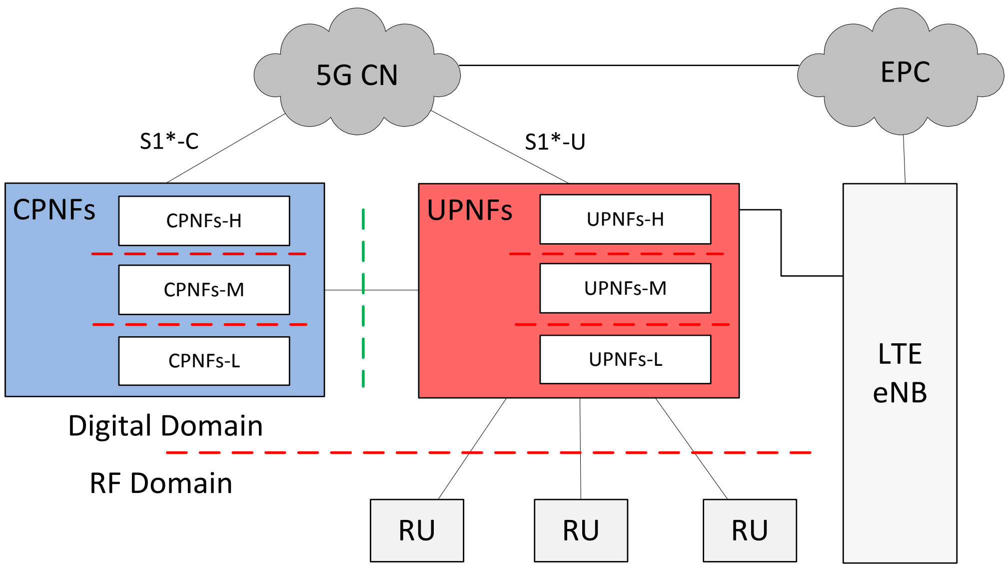


Figure 5-9. Abstract function split model considered in METIS-II.

- **Scenario 1** (stand-alone access nodes with imperfect backhaul infrastructure): In this case, the upper NFs could comprise the asynchronous layers (RRC and PDCP, and also RLC if aforementioned changes are applied), and be mapped to aggregation points, while the middle and lower NFs would comprise of the PHY and MAC layers and be mapped to the radio units. This would correspond to split options M8 or M7 in Figure 5-7.
- **Scenario 2** (centralized processing with perfect fronthaul): In this case, a large extent of centralization could be obtained. The interface between this node with centralized functions and the radio sites could be based on classical CPRI interfaces (corresponding to split M1 in Figure 5-7), or, in order to alleviate the CPRI bandwidth requirements in particular in the context of large system bandwidth and a large number of individually fed antenna elements, a split in the MAC layer such as M5 could be pursued.
- **Scenario 3** (locally centralized baseband processing): This would constitute a combination of scenarios 1 and 2: The asynchronous layers RRC and PDCP would be placed in centralized clouds, using interfaces M8 or M7 towards edge clouds. The

interfaces towards the radio units could be based on CPRI (option M1) or on a split within the MAC (option M5).

- **Scenario 4** (self-backhauling): This scenario has not yet been concluded in METIS-II.

Note that one important requirement for the 5G RAN is to provide sufficient flexibility for placement of NFs, but on the other side, the number of interfaces between the NFs, clustered on a horizontal layer structure according to the radio protocol stack, should be as small as possible. This is important in order to keep the standardization effort (as these interfaces have to be fully open-standardized) and testing effort lean (as all alternative combinations have to be tested before going into operation, which may be more stressed by multi-vendor implementations and inter-operability testing). Hence, it is not yet clear to which extent aforementioned options should be standardized or not. One possibility is to provide stage-2 specifications for these interfaces, leaving the exact degree of standardization open.

5.6 Architectural Enablers for Network Slicing

In this section, we will now elaborate on particular architecture enablers that have been identified as important to support the notion of network slices, i.e. the operation of logical networks setup for particular business cases. This topic is currently also under investigation in 3GPP. In [3GPP16-22864], it is for instance mentioned that “One key concept to achieve the goal of flexibility is network slicing. Network slicing allows the operator to provide dedicated logical networks with customer specific functionality, without losing the economies of scale of a common infrastructure”. Also the current understanding in 3GPP SA2 is that “A network slice is composed of all NFs that are required to provide the required services and network capabilities, and the resources to run these NFs.“

It should be pointed out that a network slice is expected to be related to a particular business constellation in the 5G era, i.e. the interrelation of different players such as network operator(s), over-the-top players, verticals, resellers, infrastructure owners etc., which is likely captured in the form of service level agreements (SLAs) among the players. Such a business constellation may relate to one or multiple service types (see the introduction of the main 5G services types in Section 2.1). As an example, there could be a business constellation involving a mobile network operator, a car manufacturer and an application provider, which could be related to enabling safety-critical communication between cars, as well as providing xMBB to cars for the purpose of in-car entertainment. SLAs would be setup that describe which QoS metrics are expected to be guaranteed, possibly also denoting a minimum amount of spectrum to be dedicated to this business constellation etc. In this respect, it is important to point out that **there is no one-to-one relation between network slices and services**, i.e. a single network slice may contain a mix of services, and the same type of service may also be provided via multiple network slices. As an example for the latter, a mobile network operator may provide multiple vehicular safety related services, in each case involving different parties and different SLAs, and hence using separate network slice instances for these. It is expected that the setup and configuration of network slices, and hence also the decision on how many slices to establish for which purpose, is left to the mobile network operator.

It is clear that many use cases in 5G will not succeed from a business perspective unless they share the same infrastructure with other use cases. As an example, vehicular safety could likely not be economically provided if this would require the roll-out of dedicated infrastructure. In the context of network slicing, this means that multiple slices will likely reuse infrastructure to a large extent. Within the CN, this would likely mean that multiple slices use different virtual machines for UP and CP processing, but these would be run on the same physical data centres as other slices. In the RAN, it is likely that multiple slices and the services therein would in fact be multiplexed into a common MAC, PHY and radio with other slices, though this need not always be the case (e.g. a particular slice may use dedicated spectrum and/or dedicated access nodes and hence also use a dedicated MAC/PHY etc.). Note that many variants are possible, e.g. it could make sense to have a dedicated MAC scheduler for particular slices or services. Also, even if multiple slices are multiplexed into a common MAC/PHY, certain NFs could be highly slice- or service-specific (e.g. specific H-ARQ configuration). The trend toward service multiplexing on lower layers is depicted in Figure 5-10. To have the complete picture for supporting slicing in RAN, it is important to note that these alternatives have to be considered under the light of the possibility to support different functional splits as discussed in Section 5.5.

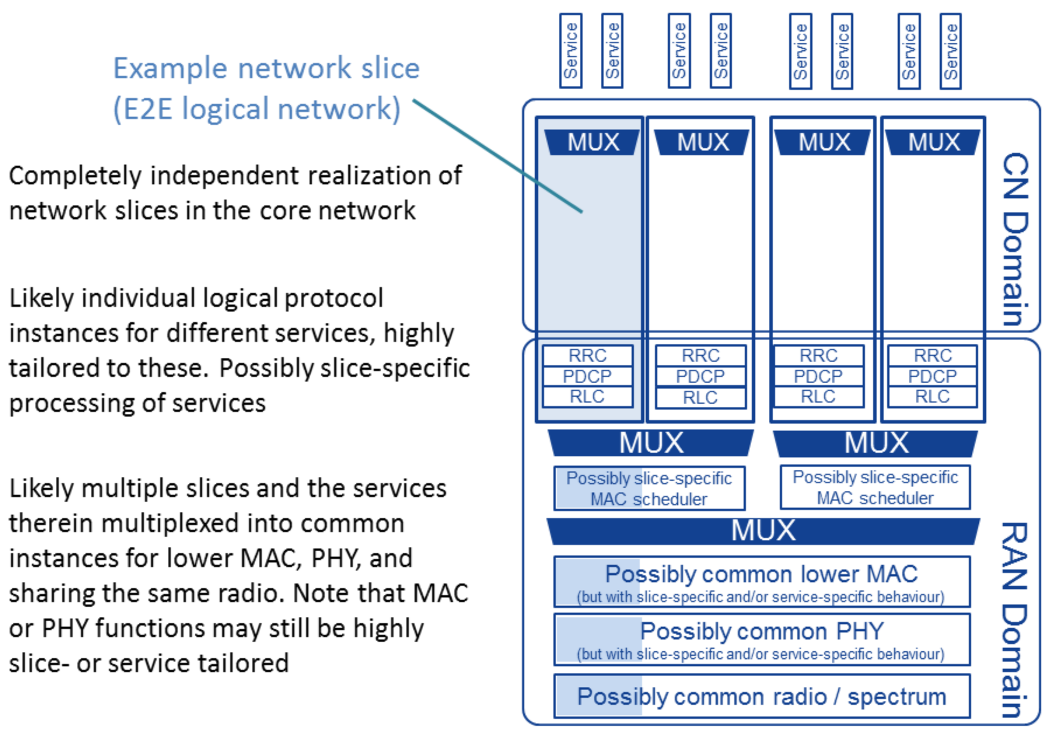


Figure 5-10.Trend towards multiplexing of slices into common lower layers, while each network slice remains a logical network from E2E perspective.

The METIS-II partners have further identified the following aspects related to network slicing:

- **Network slices** (or an abstraction thereof, such as groups of service flows) **need to be visible to the RAN**, such that NFs can take into account overall slice-specific metrics or constraints (for instance, the constraint that all services belonging to a slice may jointly only occupy a certain amount of radio resources).

- **Isolation among slices:** Slice isolation is essentially about managing network and computing resources in a way that the performance of one slice is not affected by the operation of another slice. Thus, it is required to have mechanisms to protect common channels or resources used for UEs accessing system so that congestion in one slice does not have a negative impact on another slice. Currently, 3GPP systems provide some support for protecting common control channels for extensive load from different services. These mechanisms include **Access Class Barring, Enhanced Access Barring, Service Specific Access Barring**, as well as implementation **specific admission control** etc. Moreover, another 3GPP mechanism is the **Application specific Congestion control for Data Communication (ACDC)**. As described at [3GPP16-22011], that is an access control mechanism for the operator to allow/prevent new access attempts from particular, operator-identified applications in the UE in idle mode. Further investigations are needed to **adapt these mechanisms for slice specific related functions**. As a potential technology component supporting slice isolation, METIS-II is investigating means for service prioritization, where a combination of random access channel (RACH) preambles is being used for the differentiation among high and low priority services during initial access; this combination enables minimum effect to the other slices/services [MII16-D61].

- **I**n order for the slicing concept to allow for efficient usage of common resources such as radio spectrum, radio infrastructure, and transport between the slices sharing of resources should be possible. In these cases it may also be needed to have **appropriate scheduling schemes** that will take care of fast fluctuations of traffic in different slices. Only in **special cases (**e.g., trying to fulfill ultra reliable transmission of messages with 99.999% successful transmissions at very short delays such as 1-5msec, or in cases where special regulations exist)_ it should be assumed that a slice could be assigned dedicated (static) resources (e.g. based on regulatory and/or legal requirements), since this may severely reduce the resource efficiency. Finally, since different slices require the fulfillment of different KPIs, inter-slice RRM schemes need also to be examined, since for example an attempt to increase the throughput for the UEs in a specific area may interfere with radio nodes supporting other slices (e.g., V2X slice). Please find more details on such schemes in Section 6.2.2.

- **Slice selection:** An interesting question which is still under investigation is how a device or service would be mapped to a particular slice. The simplest option would be to use pre-configured information (e.g., stored in the SIM card). This option has the benefit of being very simple and preferable in several cases (e.g., use of cheap, static sensors).

On the other hand, this option may lack the required flexibility for operators to be able to deploy more dynamically slices in area or even using different slice identifiers. Another solution is to define a default slice which all devices or services will use during their initial attachment so as to collect information about the available slices in an area. Upon their attachment at this slice, a UE has the possibility to be re-directed by a network to a tailor-cut slice for the UE. Finally, another alternative could be to use broadcast information by the network so that UE does not have to perform the attach procedure twice like in the former alternative. Also, the network will be able to dynamically update the broadcasted information about available slices in an area. Ultimately, it is not yet clear to which extent the aforementioned aspects have to be standardized or are implementation-specific.

- **Multi-slice connectivity** refers to the notion of having a single device involved in multiple slices. A potential scenario would be the following. A car may be equipped with one 5G modem that can be connected to both a V2X slice for autonomous driving and at the same time be connected to an xMBB slice so as to act as a relay for the mobile phones of the passengers. Note that it has to be clarified whether this is actually a viable option, or whether legal reasons would rather mandate the usage of individual devices for vehicular safety and xMBB. In general, multi-slice connectivity would open significant issues for the UEs especially if they would have to support more than one RRC instance (i.e. one per slice). The added complexity of such a choice is something that has to be carefully evaluated. Of course, alternative solutions would require that a UE may support multi-slice connectivity using one RRC machine, but then some of the RRC functions would require further investigations.
- **Dynamic RAN slice configuration**. The need to be forward compatible requires that the RAN allows, to some extent, the possibility of dynamically configuring or activating the slices in the RAN. This means that the network should be able to parameterize, activate or de-activate functions in a dynamic manner and based on the business driven needs of the customers. This will require the definition of appropriate management functionalities for radio reconfiguration (see Section 5.7).
- **Performance monitoring solutions** (e.g. counters, traces and KPIs) need to be aggregated per slice to verify the fulfillment of SLAs and/or properly operate the different businesses associated to different slices;
- **Configuration management**, SON etc. could be tailored, turned on/off and/or possibly configured individually for different slices.

5.7 Network Management and Orchestration[5]

The final RAN design of METIS-II is intended to fulfil NGMN's vision for the overall 5G architecture, as described in the 5G White Paper [NGM15-WP]. NGMN envisions a native SDN/NFV-based architecture that is set up on different layers covering aspects ranging from devices, (mobile/fixed) infrastructure, NFs, value enabling capabilities etc., up to all the management functions needed to orchestrate the 5G system (E2E Management and Orchestration, MANO). This approach is generally considered in the architectural description provided by the 5G PPP Working Group (WG) "Architecture" (see Figure 5-11 for a high level overview [5GARCH16-WP]). Design principles developed by METIS-II on 5G RAN are also incorporated into that architecture.

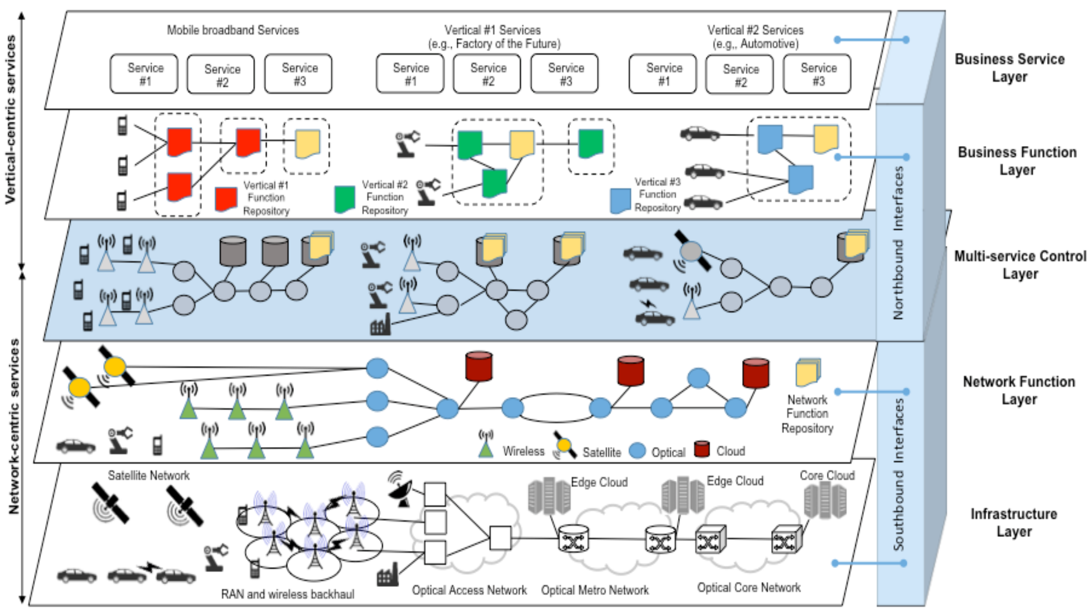


Figure 5-11. Management framework for the integration of mobile broadband and vertical services [5GARCH16-WP].

This E2E MANO is responsible for the translation of 5G use cases and business models into concrete services and network slices. Dependent on defined SLAs it determines for each slice instance and corresponding service flows, respectively, all relevant NFs, AIVs, and parameter configurations, and finally maps them onto the available 5G infrastructure (slice/service chaining) consisting of HW and SW parts including networking (radio, transport, etc.), computing and storage resources, radio frequency (RF) units and cables. The network slicing concept

5 The network management and orchestration framework is not a main research topic in METIS-II, therefore only issues with relevance to RAN design are noted in the following.

leads to high resource utilization efficiency, scalability and adaptability since each slide is designed and managed to dynamically provide the required amount of resources. Clearly, MANO requires inter-slice coordination, which has to support slice clustering and global management functionalities. The MANO framework is also needed to share the infrastructure among multiple slices and to provide efficient lifecycle management mechanisms for slice instances (i.e., deployment, operation, monitoring, and termination). It further manages scaling of the capacity of individual NFs and their geographic distribution, as well as Operations/Business Support Systems (OSS/BSS), Element Managing (EM), and SON.

The E2E approach of MANO has also to cover use cases where slices, e.g. to achieve global business availability, have to be generated across multiple domains with different administrative owners (operators, infrastructure providers, etc.). Beside the respective technical implications, novel business interfaces and charging models have to be defined for such approaches [5GEx16-WP].

MANO in 5G is generally based on principles derived within the work of the European Telecommunications Standards Institute (ETSI) Industry Specification Group (ISG) NFV [ETSI-NFV, ETSI14-MAN], but there are also aspects going beyond current specifications, especially for the RAN part, to achieve required flexibility and programmability (see e.g. [5GN15-D31]). From a long-term research perspective, network slicing could benefit from SDN and Software Defined Radio (SDR) concepts. SDN is based on the decoupling of control and data planes so as to increase in the efficient use of computational resources. The design resulting from the joint implementation of NFV and SDN concepts is expected to greatly reduce capital expenditure (CAPEX) and operational expenditure (OPEX) of future mobile systems since it limits the usage of specific hardware infrastructures, as found in traditional architectures, and optimizes the utilization efficiency of the network resources. It is important to clarify that NFV and SDN are independent and rather complementary concepts. The basic idea of SDR is defining specific radio procedures that can provide flexibility, agility, and responsiveness to be easily adapted and deployed on the virtualized baseband units, including the RF part. Specifically, SDR-enabled base stations can be operated on demand on the most appropriate AIV (e.g. related to a specific frequency band), and the virtual functional split may depend on whether they need to provide one or another service (see Figure 5-12).

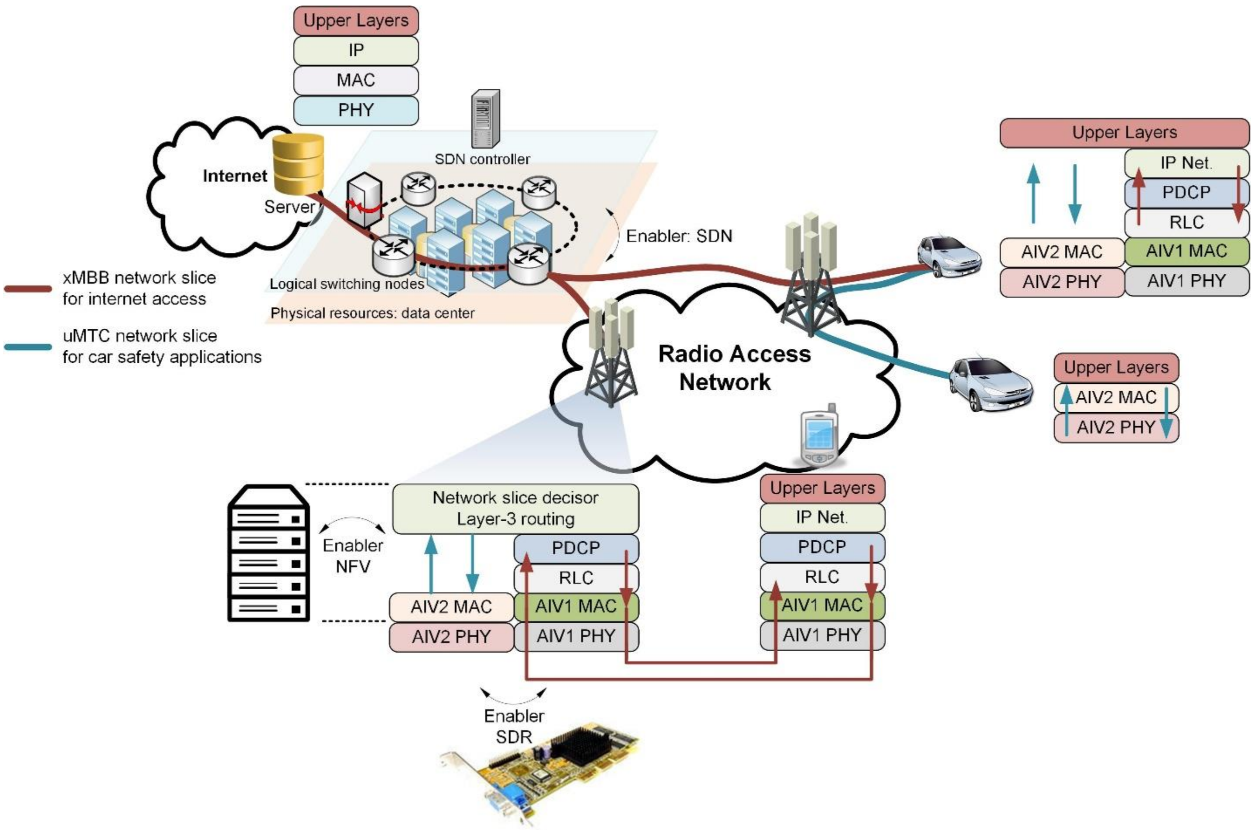


Figure 5-12. Exemplary representation of the interrelation of network slicing enablers, including SDN/NFV/SDR

There will be added complexity to introduce and handle SDN/NFV/SDR principles (e.g. interfaces/APIs, performance monitoring, security issues). Therefore, it is still an open issue how far virtualization in the RAN can be cost-efficiently introduced while achieving high performance without extensive usage of dedicated reconfigurable HW accelerators (i.e., Physical NFs in contrast to Virtual NFs), which may limit flexibility.

Especially the interrelation between logical and infrastructure resources in the RAN and their efficient and fast management and orchestration via hypervisor principles has to be further evaluated (e.g. re-adjustment and/or re-configuration of NF resources in case of overload or fault situations). Novel management schemes for AIV reconfiguration (e.g. changing the properties of an AIV such as the carrier bandwidth, frame numerology etc. on the order of hours) may be adopted for more efficient use, to adapt the network to the dynamic behavior of the traffic and to globally maximize the capacity. Accordingly, the AIV reconfiguration can shape the overall AI landscape and implies a modification to the available set of resources on which RM schemes will be operating. Therefore, the functions within agile RM framework, as detailed in Section 6.2, need to take into account such AIV reconfigurations. In fact, given a cell set in a certain area, the traffic of different slices or services e.g. on a specified AIV may change from one sub-area to the other according to the day period (e.g., in a stadium before, during and after an event such as a football match or a concert). It often happens that some areas may be congested (with high blocking percentages) in some particular zones (the so-called hot-spots as the stadium) in which the traffic is more consistent, while surrounding cells are less loaded or characterized by low blocking percentages.

The AIV reconfiguration management functionality spans different AIVs and managing and controlling the nodes inside the NW, with the goal of self-adapting towards an optimal mix of supported AIVs and frequency bands. This function could act on the basis of some input parameters, such as the available resources (spectrum and HW), the traffic demand, the capabilities of the UEs within the NW (supported AIVs, frequency bands, etc.), the requested services (e.g., bandwidth and QoS), etc. In addition, this functionality could exploit a collaborative AIV reconfiguration management scheme, where the decision making functions are shared among different NW nodes.

On these basis, two main different typologies of reconfiguration according to the specific contexts can be performed: an intra-system reconfiguration that involves only one single AIV and/or an inter-system reconfiguration that involves two or more different AIVs. For example, an intra-system reconfiguration could be necessary when the traffic on a specified AIV drastically changes from one sub-area to the other (e.g. some cells may be congested while the surrounding ones are not). On the other side, the inter-system reconfiguration could be needed when different traffic loads are experienced by each AIV, in order to increase the percentage of radio resources devoted to the over-loaded system while decreasing the ones used by the others (supposed under-loaded). It should be noted that intra-system and inter-system reconfigurations can be simultaneously performed. Figure 5-13 depicts a generic scenario that includes both types of reconfigurations and that implicates modifications to both hardware and radio resources for the involved AIVs. Please note once again that the term AIV reconfiguration here refers to the change of substantial parameters of AIVs (e.g. change of channel bandwidth, change of numerology, activation of a novel AIV in a different frequency band etc.), which is expected to happen on a slower time scale (e.g. on the order of hours). A fast activation / deactivation of access points and mapping of radio resources within a given AIV configuration or over multi-AIVs are handled in the context of the agile RM framework presented in Section 6.2.

As anticipated, in order to perform such network reconfigurations, an appropriate AIV reconfiguration management functionality to span different AIVs, manage and control the nodes inside the network, in order to self-adapt towards an optimal mix of supported AIVs and frequency bands, need to be introduced at the logical CP level. From a high level perspective, such AIV reconfiguration management functionality is devoted to perform the following actions:

- Monitor periodically the current activity status of the cells (for each supported AIV), for example, in terms of measurement of the number of the requests and rejects (if any) from the different systems;

- Checking conditions of the associated cells and, depending on the results of the check, execute a reconfiguration algorithm to identify which nodes need to be reconfigured, e.g. with the aim to adapt the percentages of processing resources devoted to each supported AI/AIV and to dynamically shape the active radio resources to the behaviour of the traffic;
- Control the network reconfiguration by sending appropriate reconfiguration commands to the reconfigurable nodes in order to perform the appropriate actions (for example to activate/deactivate processing resources and/or radio resources, such as frequency carriers, for each supported AIV).

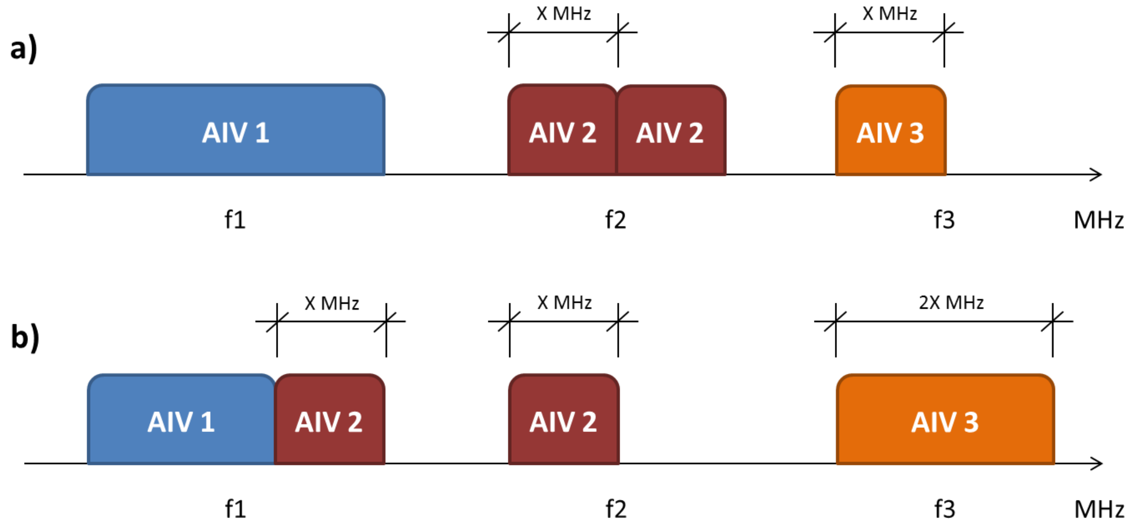


Figure 5-13: Examples of AIV reconfigurations that implicate modifications to both hardware and radio resources.

The AIV reconfiguration management functionality can be configured to perform the above mentioned actions periodically, starting, for example, from an initial condition set by the network operator and depending on network planning parameters. For example, a monitoring period of length *T* can be defined to collect the cells status information and the end of which the appropriate checks to evaluate and act a reconfiguration of the network are performed. It has to be noted that performing too many reconfigurations in a short time (e.g. small values of *T*, such as 30 minutes) could require to send too many signaling messages for configuring and reconfiguring the nodes. This could lead the system to an unstable situation due to too fast modifications of the planning. On the other hand, performing too few reconfigurations in a long time (e.g. high values of *T*, such as multiple hours) could not be much effective, since the time elapsed between the need of a reconfiguration and the time of the reconfiguration itself could be too long. Such factors should be taken into account in order when setting the value of *T*.

6 Functional Design Considerations

In this chapter, a summary of various functional design considerations of METIS-II is provided. Initially, general design considerations are listed in Section 6.1, which are expected to have a strong implication on the overall 5G RAN design. Section 6.2 then ventures into more detailed considerations on the functional design for traffic steering and RM in 5G, and Section 6.3. provides details on design considerations related to initial access and mobility. The chapter is concluded with summarizing the key paradigm changes inherent to the depicted functional design considerations, the main differences with respect to LTE-A, and the implications on the overall 5G RAN design.

6.1 General Considerations

6.1.1 Service-Tailored Network Functions in 5G

Flexibility and configurability will be key RAN characteristics to support the diverse services and related requirements in the forthcoming 5G ecosystem. This may be realized by a protocol architecture supporting a service-specific selection of network functions (NFs) and service-tailored optimizations. At the same time, as stated already in Section 4.2, METIS-II envisions that the overall 5G AI should ideally be characterized by a large extent of **protocol harmonization across the AIVs used for different bands, services and cell types.**

In Table 6-1, we provide some specific examples of NFs that could be tailored to specific service needs in 5G, taking some input from [5GARCH16-WP]. In general, there is the common understanding that specific services will likely reuse the same functionalities as other services for a large portion of the protocol stack, differing only for a smaller number of functionalities. For instance, it may be possible to use a flexible AI numerology and, depending on the service needs, different coding strategies, MIMO modes and framing structures optimized for throughput, delay, or reliability. The upper layer packetization, however, may still be the same for a wide range of services, allowing to reuse the same software implementation. Note also that a number of CP functions may require a different configuration of protocols for different slices and possibly - in some extreme cases - the use of new protocols.

Table 6-1. Potential service-specific flavors of network functions, partially taking input from the 5G Architecture WG White Paper [5GARCH16-WP].

Type of network function		Possible service-specific flavor
General connectivity	Connectivity model	e.g., bearer-based (for high throughput services), or connection-less (for internet of things, IoT).
	Multi-Connectivity	Multi-connectivity at different network layers (micro/macro), technologies (WiFi/LTE), spectrum (sub-6 GHz/mmWave), user plane layers (MAC/RLC/PDCP) depending on service, deployment and AIV (see, e.g., Section 6.2.1 and [MII16-D51, MII16-D61]).
Spectrum Access		Service-dependent operation in licensed, unlicensed, or license-assisted spectrum, or time-frequency multiplexed in common spectrum (see, e.g., the extended notion of resources and specific considerations in [MII16-D51]).
Advanced SON schemes		Support of the dynamic densification through agile RAN schemes, e.g., Nomadic Nodes (see, e.g., interference management based on dynamic radio topology in [MII16-D51]).
RRC related	Mobility	No (metering), local (enterprises), in groups (trains), very high speed (cars/trains/aircraft), on demand/forward (tracking sensors) or always/backward (pedestrian broadband) handover.
	Cell discovery	Sub-6 GHz MIMO (broadcast), massive MIMO mmWave (sub-6 GHz assisted), small cells in ultra-dense networks (via macro coverage layer) cell discovery.
PDCP		Potential service-specific omitting of header compression and ciphering even for user plane traffic.
RLC		Potential service specific unacknowledged mode only (e.g. sensor) or acknowledged mode only (e.g. mission-critical services), or transparent mode.
MAC / PHY	Carrier Aggregation	Carrier aggregation may not be needed in each scenario as it also impacts battery consumption; it could further include very distinct spectrum.

	Multi-Cell Cooperation	Service, load, deployment and channel-dependent tight cooperation (symbol-synchronized operation, RNTIs/scrambling/CSI-RS/scheduling/precoding coordination up to joint Tx/Rx CoMP) or loose cooperation (ICIC) (for specific considerations see, e.g., Section 6.2 and [MII16-D51]).
	Scheduling	Service specific scheduling schemes, as for instance semi-persistent scheduling on sidelinks using geo-location information to improve V2X communication performance.
	RACH	Service specific RACH schemes where priorities can be introduced (please note a specific proposal for RACH service prioritization, which is described further in [MII16-D61]). Also, grant free schemes can be considered for services to minimize the establishment of signaling channels or the transmission of emergency data.
	H-ARQ	Optimized for spectral efficiency (massive broadband) coverage (sensor, IoT), reliability (mission critical services) or latency (tactile Internet).
	Coding	Block codes for short (sensor) transmissions, turbo-codes for high throughput.

6.1.2 Beam-centric Design

To support the long-term traffic demands and efficiently enable multi-Gb/s data rates, 5G operation will very likely not be limited to frequencies below 6 GHz, as currently used for LTE-A, but comprise bands up to 100 GHz. Compared to the bands used for LTE-A, much more challenging propagation conditions exist for higher bands, such as lower diffraction and higher outdoor / indoor penetration losses, meaning that signals have less ability to propagate around corners and penetrate walls. In addition, atmospheric/rain attenuation and higher body losses could also contribute to making the coverage of higher frequency AIVs spotty. An extensive usage of beamforming, where multiple antenna elements are used to form narrow beams, will be essential in 5G to overcome these propagation challenges, and will be facilitated by the fact that higher frequencies allow for larger antenna arrays at reasonable form factor.

The 5G AI is envisioned to support a massive use of beamforming especially in higher frequencies, in what is called a beam-centric design. In a beam-centric design, both data and control channels (at least dedicated channels) would need to be designed so that they can be beamformed via different beamforming architectures (analog, digital and hybrid). When it comes to common channels (such as those carrying system information, paging information and synchronization signals), additional challenges would exist, and in some deployments the design of beam finding and beam sweeping would need to be supported. In that case, the UE would synchronize with a beam, obtain system information associated to a beam (or a group of beams) and access the medium accordingly.

A beam-centric design may have substantial implications on how RM is efficiently and effectively performed. One example is the implication on the dynamic traffic steering, which aims at a fast re-routing of the data flows to the appropriate links (see Section 6.2.1). In this regard, the radio link feedback could include the beam characteristics along with the beam ID in order to determine the reliability and capacity of the links associated with such beam. A beam-centric design may also have an impact on interference management schemes. For example, in case Frequency Shift Keying and Quadrature Amplitude Modulation (FQAM) are applied to reduce the effect of interference on cell-edge users, it needs to be decided on which beam such FQAM modulation is to be applied. Another example is the case of utilizing dynamic TDD in dense small cell deployments. In this case, a beam-centric design can be exploited to alleviate the impact of same-entity interference (i.e., user-to-user and BS-to-BS), which can be severe. Furthermore, a beam-centric design requires an extended UE measurement context, as discussed further in Section 6.2.5.

A beam-centric design may also affect the way mobility is handled. First of all, reference signals used for neighbor link measurements should be designed so that they can be beamformed. Similarly, the switching upon link degradation should be between beams rather than cells (even if in some cases a beam may be as wide as a cell sector in LTE-A). Another difference relates to the time to perform link measurements. For higher frequencies, the signal-to-interference and noise ratio (SINR) can drop quite quickly due to the mentioned propagation conditions, so that the long term filtering applied in LTE-A may lead to too many RLFs and/or mobility failures. Similar challenges apply to the design of multi-connectivity solutions where multiple links are represented by beams. It needs to be investigated to which extent a beam-centric design changes the cell-based paradigm. The beam-centric design could be used to support the concept of virtual or user-centric cells, also known as cloud cells. More details can be found in [MII16-D61].

6.1.3 Lean and Future-Proof Design

Further, it is being investigated how to obtain a lean and future-proof design that maximizes energy efficiency, reduces the amount of interference generated by common signals, and better supports beam-centric communication and the possible introduction of future control signals. A first step to achieve this is to minimize broadcasted signals by using dedicated transmission whenever possible, as shown in Figure 6-1 a). System information distribution should be designed such that only the most fundamental information needed for connection setup is transmitted in common channels for inactive UEs. One may also strive to avoid transmitting, e.g., reference signals over the entire bandwidth, but instead use self-contained transmissions as shown in Figure 6-1 b), where reference signals are transmitted jointly with the payload, minimizing overhead and interference, and being better suited for beam-based transmission and future-proofness.

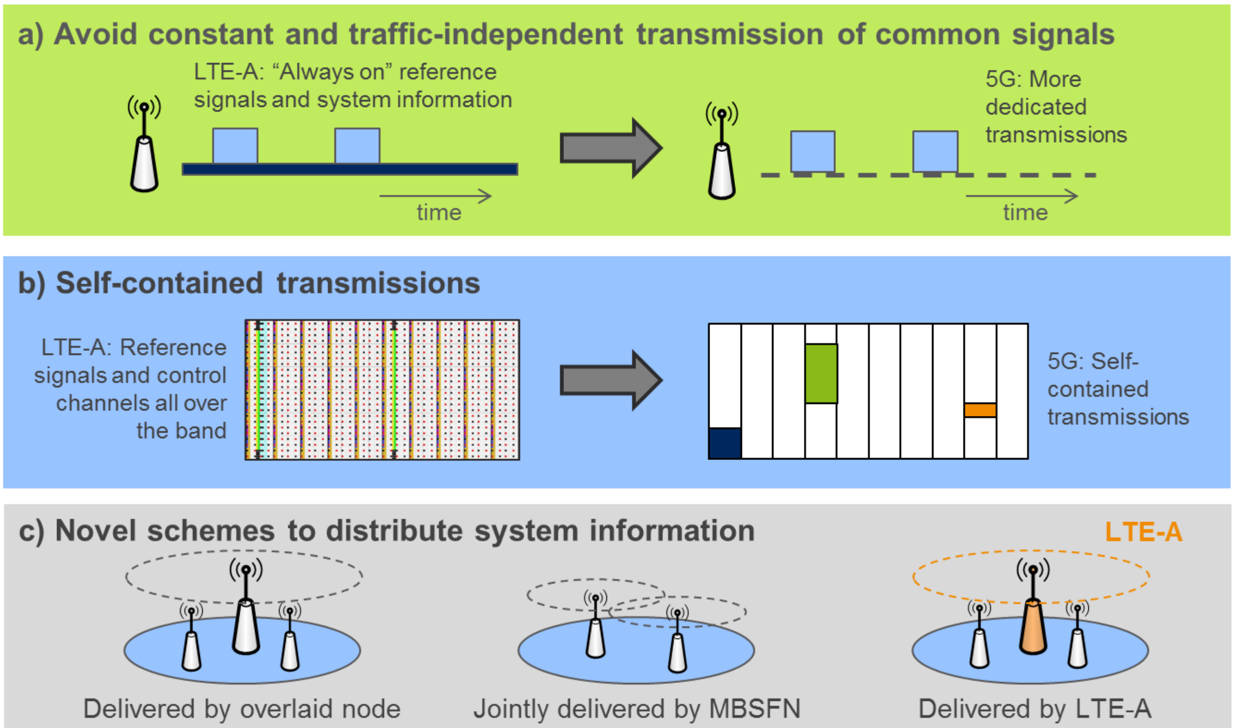


Figure 6-1. Means to obtain a lean and future-proof design; MBSFN refers to Multicast-broadcast single-frequency network.

One example of this is the principle of in-resource control signaling, which can be used together with a flexible frame structure with variable TTI size, to dynamically signal user scheduling grants in the fraction of resources that are also used for data [MII16-D51]. This is different from current LTE-A, in which a strict and periodic time-division separation between control and data exists, with e.g. physical layer downlink control signaling being transmitted over the full system bandwidth in the first OFDM symbol of every TTI and with a fixed duration of 1 ms. Another potential improvement is a design that allows only a subset of access nodes to use common channels for system access, e.g. by only letting some nodes transmit system information, or by applying a CP/UP split among cells, as shown in Figure 6-1 c). See more details in [MII16-D61].

6.1.4 Energy Efficient Design and RAN Moderation in 5G

While it is foreseen that traffic will increase massively in 5G networks, a key element to allow sustainability of future systems is that the absolute energy consumption in cellular networks should not increase as compared to today. In essence, this means that the overall energy efficiency (e.g. in bits/Joule) will have to increase at a similar or even higher pace than the mobile data traffic itself.

The topic of energy consumption in mobile networks has been already widely studied in recent years [EARTH, 5GREEN], having led to the availability of new transmission nodes with more efficient power consumption profiles, that are able to scale better their power usage based on the actual amount of traffic that is served [MII16-D51]. As explained in Section 6.1.3, the lean design of the 5G common channels is one important enabler for this. Also, mechanisms that enable fast on/off switching of nodes already exist, allowing to further reduce the energy consumption of a node when it is not transmitting, exploiting a “sleep mode” or “lock” state (see also [WZZ15]).

The higher degree of flexibility and of coordination that will be available in 5G networks allow to increase the saving offered by this kind of mechanism. For example, in [MII16-D51], an approach is investigated that exploits fast sleep mechanisms through a centralized scheduler, which leverages on coordination techniques such as Joint Transmission and Dynamic Point Selection / Blanking to further reduce energy consumption when traffic is below its peak. This approach can work on a very short time scale, in the order of milliseconds, working on instantaneous fluctuation of the overall traffic. In [MII16-D51] it is further shown that thanks to a coordinated management of the available transmission nodes it is possible to trade-off the extra capacity present in the network, and achieve higher energy savings when this additional capacity is not needed. It was evaluated that a power consumption reduction of up to ~50% could be reached, compared to a situation where no coordination between the transmitting nodes is present. This coordinated approach is in particular able to reduce the power consumption depending on the dynamic portion of the energy consumption profile of base stations, i.e. the portion that depends on the amount of radio resources used for transmission. On the other hand, the static portion of the energy consumption model, the one that is always present, even when no traffic is served, cannot be reduced in this way. That portion represents a fundamental limit on how much energy consumption can be reduced, as was already highlighted in the EARTH project. In that sense, there is a limit on the effectiveness that can be reached exploiting only sleep modes, since there has to be some signals that are always on to make the CP of the system work properly. This issue can be solved with novel approaches based on a lean design of the network.

Energy efficiency in the context of joint transmission and reception for dynamic TDD is also investigated as part of [MII16-D51]. Dynamic TDD is considered a promising solution to cope with fast-varying traffic, especially in ultra-dense small cell deployments where traffic is driven by only a few active UEs. Clearly, throughput and energy efficiency are in a trade-off relationship, as any smallest increase in throughput requires powering on sleeping BSs.

It is further shown in [MII16-D51] that a network energy efficiency utility can provide comparatively good performance increase in terms of average UE throughput for an indoor environment, while at the same time reducing energy consumption by keeping BSs that provide little to no gain in sleep mode. The largest performance gains are observed at low traffic load, which is when the number of BSs that can be selected for the joint transmission and reception is greatest, and the generated interference to other UEs is limited.

6.1.5 Native Relaying, Self-backhauling and D2D Support in 5G

As mentioned in Section 3.1, a key design requirement of the 5G system is the native support of relaying, self-backhauling and D2D, as opposed to legacy systems like LTE-A, where these features are either introduced as an extension to the original design or have not yet been introduced. Such add-on approach in many cases naturally involve compromises w.r.t. a potentially better design. METIS-II is investigating the following communication scenarios that are related to relaying, self-backhauling and D2D:

- **Grouping of devices in proximity with similar communication needs**. Multiple devices in proximity may be grouped together based on their mobility, similar service and communication characteristics (e.g., data to be transmitted, packet delay requirements). Devices in the same group may use unicast D2D communication or one-to-many / one-to-all D2D communication. Besides, one of the group members, based on certain criteria such as power or processing capabilities, may be selected as a group head or cluster head. This group head may then use the PC5* interface with its directly connected devices for D2D discovery and / or communication for collecting aggregating CP messages (e.g., RACH requests). The intra-group communication may take place either via a different interface e.g. IEEE 802.15 / Zigbee or IEEE 802.11 or via D2D communication over a cellular AIV. The group head may use the Uu interface for communication with the 5G-RAN. Devices in the same group can jointly access a cellular system instead of individually doing so. This communication scenario could be solved by operating cluster heads as self-backhauled access nodes (i.e. such that other devices in proximity would perceive these as infrastructure nodes and connect to them accordingly), or by keeping cluster heads and other devices on the same hierarchy level and using peer-to-peer (or multicast) D2D communication among these.
- **Deep coverage extension for mMTC services**. Deep coverage refers to the case where mMTC devices are deployed in locations where they experience challenging radio propagation conditions with a large penetration loss, etc. Often such devices may be sensor nodes with no mobility. In such scenarios, certain mMTC UEs with decent

radio conditions may be pre-configured or dynamically selected to act as relay UEs. This helps to improve power consumption as well as increase overall coverage area.

- **D2D communication in the context of mobility**. Mobile devices may discover other devices of interest which are in their proximity, and establish D2D communication. A pair of D2D devices in proximity and with ongoing communication can be within the same base station coverage but due to their mobility, they may move out of coverage of one base station and in-coverage of another base station. Here, the focus is on group mobility and UE mobility under different inter-AIVs or intra-AIV scenarios.
- **Wireless self-backhauling in very dense 5G deployments**. It is commonly understood that for very dense small cell deployments in 5G it may be economically unfeasible to establish wired backhaul for each access node. Hence, wireless in-band backhauling should be natively supported, enabling access nodes without wired backhaul to autonomously establish backhaul links using the same cellular technology as the wireless access.

Since the enablers for relaying, self-backhauling and D2D have many commonalities, these are here treated in one sub-section. METIS-II has been working on developing following solutions:

- **Channel sounding among pairs of devices**. One key requirement in the context of D2D communications is the need to estimate links between devices. The METIS-II assumption is here that this should be done based on a reuse of the same sounding reference signals (SRS) that are also used for the cellular uplink. This is different to the LTE Rel. 12 approach, where dedicated signals are used for device discovery. A key challenge is then that a device can of course only send its own SRS or receive the SRS transmitted from another device at the same time. Hence, it is required to design SRS muting patterns such that devices can estimate the links to other devices in proximity over time.
- **Control signaling among devices**. Another difficulty is how to enable control signaling between directly communicating devices (e.g. ACK / NACK, channel quality indicator, CQI, feedback etc.), in particular if this is expected to build upon the same control channels as designed for cellular communications. If for instance there are certain signals foreseen for uplink control signaling, and others for downlink control signaling, then if the uplink control signals are reused for the control signaling between D2D pairs, a device can only transmit uplink control signals or receive these from another device, but not both at the same time. Solutions are here to either again apply a muting pattern to the control signals, as in the case of channel sounding, or to relay control signals via an infrastructure node.
- **Cooperative D2D communication** is when D2D pairs are utilized as relays to facilitate the transmission between a cellular user (CU) and its base-station (BS) to improve spectrum efficiency. In this case, the PC5* interface is enhanced to support unicast D2D communication and / or one-to-many / one-to-all D2D communication among pairs of

devices, where one of these devices can be the source (DT or D2D Transmitter), while other devices can be the destination (DR or D2D Receiver). Besides, such D2D devices facilitate cellular user transmission by acting as relay devices. A cooperative communication scheme as proposed here enables 5G RAN to dynamically allow cooperative D2D mode selection and communication, while at the same time ensuring interference mitigation e.g. in case of simultaneous D2D and CU to BS communication over the shared radio resources, etc. More precisely, three types of cooperation are considered, namely overlay, underlay and hybrid cooperation [MII16-D61].

- **D2D discovery and communication**. In a D2D discovery procedure, a BS performs the D2D pairing algorithm by analyzing dynamically collected context information. After receiving BS confirmation, a relay UE sends a discovery reply message together with a reference signal to the remote UE. The remote UE may respond with an ACK or NACK based on its calculated RSRP of the D2D link. In a D2D communication procedure, a remote UE transmits its data packet to the corresponding relay UE with which it has been paired in the D2D discovery procedure. A random access procedure and D2D link configuration procedure might occur in this step based on whether the uplink traffic from the remote UE is periodic or not. Afterwards, the relay UE will forward the successfully received packets to the serving BS. The proposed D2D discovery and D2D communication procedures are supported in idle as well as connected inactive state. It is only necessary for a relay UE to enter RRC active state if the relay UE needs to forward the result of a discovery procedure or a data packet of a remote UE to the BS. The required configuration information to support D2D operation is carried by D2D link system information blocks and downlink control information from physical downlink control channel (PDCCH). Moreover, when paired D2D UEs stay in connected inactive state, certain context information related to the D2D link can be kept in both relay UE and remote UE(s), in order to reduce both signaling load and power consumption.
- **Self-backhauling**. A basic functional requirement for self-backhauling is the need to be able to align the transmissions on backhaul and access links from the perspective of a self-backhauled entity. If, for instance, a self-backhauled node uses one transceiver for both access and backhaul, it must be possible to multiplex the control signaling on the backhaul and access links in time, meaning that both the backhaul and access link must be able to be configured to use certain muting patterns w.r.t. control signaling. If a self-backhauled node has separate transceivers available for backhaul and access, this half-duplex constraint would of course be relaxed, but it may still be necessary that the node synchronizes the usage of transmission and reception on the backhaul and access links (i.e. it may not transmit control signals on the backhaul link while it is receiving control signals on the access link, due to potentially too large cross-interference).

6.2 Functions related to Agile Traffic Steering and Resource Management

The diverse set of requirements from the wide range of services and the need for a sustainable 5G system (e.g., in terms of low energy consumption, high flexibility, and ability to support new businesses) necessitate efficient and effective RM schemes that work on an extended realm of resources [MII16-D51]. To this end, METIS-II is developing an agile RM framework that holistically considers the novel and differentiating aspects of 5G systems with respect to previous generations of mobile communication standards, specifically in terms of diverse and challenging services and use cases, existence of multiple AIVs, dynamic radio topologies (e.g., nomadic nodes), and novel communication modes (e.g., D2D). In what follows, various functional considerations within this framework are briefly highlighted. Details can be found in [MII16-D51].

6.2.1 Multi-AIV Resource Mapping

To ensure that the QoS requirements of the traffic flows are fulfilled considering the relative unpredictability of the radio links (especially on the higher frequency bands), the conventional traffic steering mechanisms need to be extended towards a multi-AIV dynamic traffic steering framework, which is not only relying on hard handovers but exploiting multi-connectivity and enabling traffic flow adaptation on a faster and possibly synchronous time scale. Such a framework needs to take real-time feedback from the multiple AIVs currently serving the UE, in order to adjust the traffic flows on a fast time scale.

One possible implementation of such framework is to apply hierarchical control functionalities in the RAN. In particular, as illustrated in Figure 6-2 (left), the ‘outer loop’ RAN traffic steering functionality (AN-O) has a global 5G RAN view, in order to enable better traffic steering within each AIV. The QoS policies are sent from the CN to the AN-O layer, where the traffic aggregation is assumed to happen. Thus, the dynamic traffic steering framework is then implemented in the outer layer, in order to do a fast traffic re-routing to the various AIVs in the ‘inner loop’ RAN (AN-I) layer. The AN-I layer is operated with an optimal amount of active links engaged in multi-connectivity with the 5G UE, in order to achieve the QoS targets of the service flows. Here, the steering is assumed to happen based on real-time feedback (per-TTI or periodically over a few TTIs) from the AN-I layers. Due to the relative unreliability of the involved links, the feedback is required real-time, in order to do a fast traffic rerouting in case a link failure is detected. Current LTE radio link failure detection and recovery mechanisms would take several seconds in order to re-establish the radio bearer, and since this is unacceptable for high-priority, high-reliability traffic, the dynamic traffic steering framework will ensure that the QoS policies received from the CN are successfully enforced by avoiding radio resource reservation due to the RLFs.

In terms of the physical deployment scenarios as discussed in Section 5.5.1, multi-AIV dynamic traffic steering concepts are especially suitable for cloud-based deployments. In this regard, the potential options for a flexible protocol split between the 5G radio frontend and centralized processing are shown in Figure 6-2, based on [3GPP16-160043] with an open xHaul* assumed to be present between the radio access point (local AN-I layer) and the cloud RAN (assumed to be the local AN-O layer). Here, split A could be considered similar to the LTE-A dual connectivity feature. Option-B considers the RLC PDUs being transported over the xHaul* interface to multiple AN-I nodes and delivered to the UE which then does the combining. Option-C could be considered similar to the carrier aggregation feature in LTE, where the MAC PDUs are delivered to the UE using multiple LTE RRHs. From the dynamic traffic steering framework perspective, if the split is done at the MAC layer (option C) or lower, then there are potentially no new impacts perceived on the xHaul*, since the AN-O layer would be receiving real-time feedback for scheduling the physical resources. Yet, if the split is done at a higher layer, (options A and B), with fast traffic re-routing done over the multiple AIVs, then new RAN measurement information elements should be defined to be transported over the xHaul*, in order to enable the envisioned traffic steering. The feedback in this case should be optimized to avoid any significant additional signalling load over the xHaul* interface. In a nutshell, in 5G, due to the stringent service requirements and due to the relative unpredictability of the radio links (especially on the higher frequency bands), it is desirable that traffic steering is enforced in the lower layers of the protocol stack, for e.g., in the RLC or MAC layers.

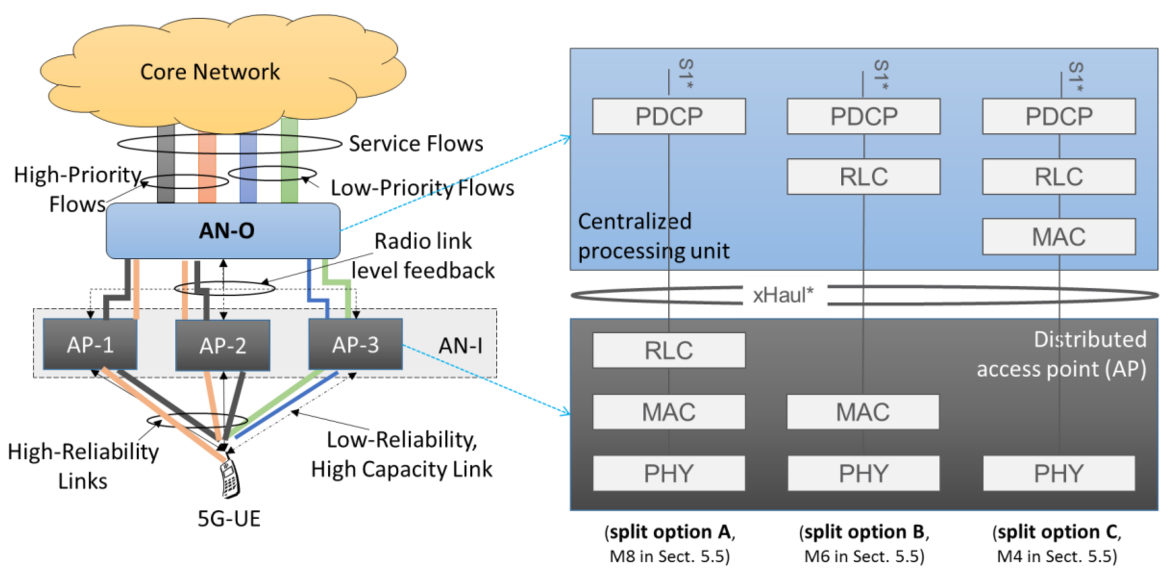


Figure 6-2. Service flow delivery mechanism considered using the proposed dynamic traffic steering framework (left), and functionalities covered by AN-O and AN-I for different example function split options (right) [MII16-D51].

It is expected that the AN-O can provide a unified and aggregated view of the various AIVs and resources at its disposal if it is implemented as an AIV-agnostic convergence or abstraction layer. A key question in this respect is which traffic steering and RM functionalities can be designed as AIV-agnostic, and which should be AIV-specific, and at which level in the protocol stack the transition should happen. Here, AIV-agnostic implies that the corresponding RM functionality could remain agnostic to the design of the physical layer design of the AIVs that are involved and, thus, can operate in an AIV-overarching manner. Furthermore, the abstraction models and RM framework should also facilitate achieving an edgeless user experience in dynamic topology settings. A convergence or abstraction layer can provide a unified and aggregated view of the various AIVs and resources at disposal, which shall be analyzed together with the associated architecture and interface implications.

Some MAC or synchronous control functions might be most suitably implemented as AIV-agnostic or AIV-overarching, such as logical channel prioritization, (de-)multiplexing of logical channels, queue management, feedback control to higher layers, AIV-specific configuration control functions, and an adaptation layer towards higher layers (e.g., to reduce dependencies of RLC parameters on radio). The scheduler, however, can be designed with different degrees of AIV-agnostic versus AIV-specific functionalities. Two examples are described in the sequel and shown in Figure 6-3.

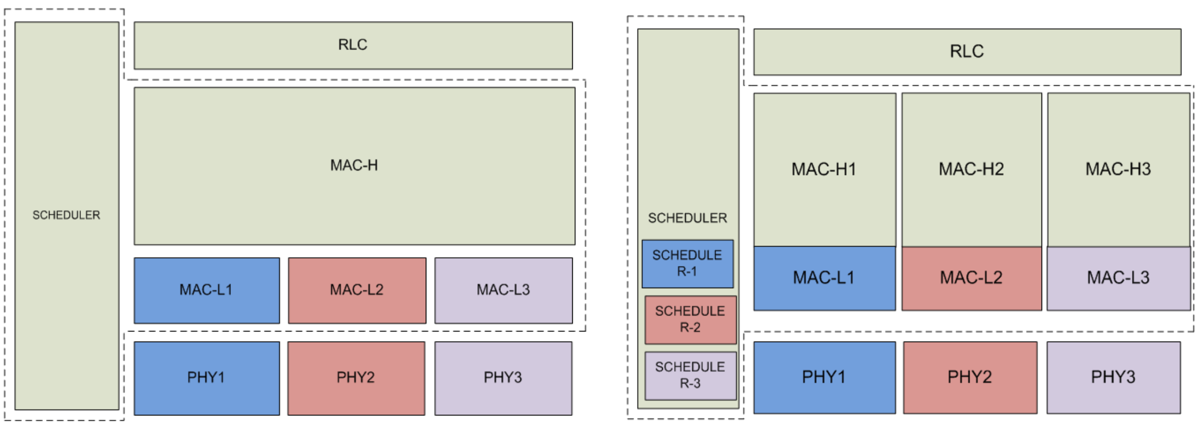


Figure 6-3. Example of AIV-agnostic RM with integrated (left) or coordinated MAC instances for multiple AIVs (right) [MII16-D51].

Example 1 of AIV-agnostic RM: Integrated MAC layer for multiple AIVs

In this option, a single MAC entity would handle different AIVs with specific MAC layer functions or sublayers which can handle the radio-aware part of each. The common or AIV-agnostic layer of the MAC performs the controlling role of configuring different radio specific entities. The scheduler in this view is common across all radio interfaces and configures parameters across all of them. MAC-H (High) as a common layer is envisioned to encapsulate the interface with higher layers and the common control/coordination functions.

Example 2 of AIV-agnostic RM: Coordinated MAC instances for multiple AIVs

In this logical view, separate instances of radio agnostic MAC are instantiated for each AIV. There are separate radio schedulers for each of the AIVs coordinated by a central scheduler entity. This provides the flexibility to have independently optimized scheduling algorithms for individual AIVs or use cases. Each radio scheduler could be handling a separate AIV or a specific traffic type. Individual MAC instances would coordinate to provide a unified framework, with individual schedulers coordinating or under a joint overall scheduler/coordinator.

It is worth noting that that it is still being researched to which extent the above concepts are purely implementation-specific, or whether these would require standardization, for instance, due to the need to introduce changes on the Uu interface.

On this basis, a flexible scheduling framework that is able to simultaneously accommodate users with very different service requirements is investigated [MII16-D51]. The design aims at full flexibility in the sense that it does not require separation and reservation of resources for different services, adapting dynamically to the traffic demands with maximum resource efficiency. For instance, the targeted flexibility can be obtained via a flexible frame structure with variable TTI size support. The developed aspects shall natively support D2D and self-backhauling, as detailed in Section 6.1.5.

The framework of multi-AIV resource mapping can incorporate not only novel AIVs, but also the interworking between legacy and novel AIVs. In this regard, one example implementation is provided considering interworking between LTE and novel 5G AIVs.

Particularly in early 5G deployments, when novel AIVs may not yet be able to provide full coverage, a tighter interworking with LTE and novel AIVs may be crucial in order to ensure ultra-high reliability and extreme bit rates in a 5G system [MII16-D51]. One possible mechanism to enable tighter interworking is LTE Rel. 12 dual connectivity [3GPP13-36842], which typically operates on a slow time scale, i.e., in an asynchronous manner. Nevertheless, considering the stringent 5G requirements, e.g., reliability and short delay, there is an urge to enable such tight interworking mechanisms on a faster time-scale, i.e., in a synchronous manner. In the following, two concepts, namely, fast UP switching and dual connectivity are highlighted. Both concepts may benefit from possibly new UE measurements per AIV (LTE and 5G) in order to make an optimal scheduling decision, preferably on milliseconds basis if the backhaul over X2* allows for this. One possible way to achieve this is to standardize new UE measurements similar to normal LTE handover measurements but on a faster time-scale. Moreover, adding and deleting a new CP connection in dual connectivity to a user must be very fast and lightweight in terms of signaling to support ultra-reliability requirements.

The first concept is a fast UP switch at the (common) PDCP layer, see Figure 6-4. At first, it is assumed that the control plane is using “dual connectivity” with LTE and 5G, while the UP is switched at PDCP level to either LTE or 5G. If the CP is connected to both the LTE node and the 5G node, no signaling is required and the UP switch may be almost instantaneous. The fast

UP switch can be based on normal handover measurements such as RSRP. However, possibly new kind of UE measurements is necessary to efficiently optimize the performance of the fast UP switch. A drawback of having multiple flows of the CP is the increased overhead.

The second concept relates to the case when both UP and CP are connected to LTE and 5G (similar to "dual connectivity" in LTE) and the UP data is aggregated (or split) at PDCP layer, see Figure 6-4. It is worth noting that an alternative to the dual connectivity solution is to use the MAC layer for aggregation, as in carrier aggregation for LTE. In this case, the scheduler can then use resources in an optimal way based on the measurement information about all carriers (i.e., both LTE and 5G carriers). The measurements and signaling to support this should also be possible to develop for the dual connectivity solution (still using PDCP as aggregation / split layer).

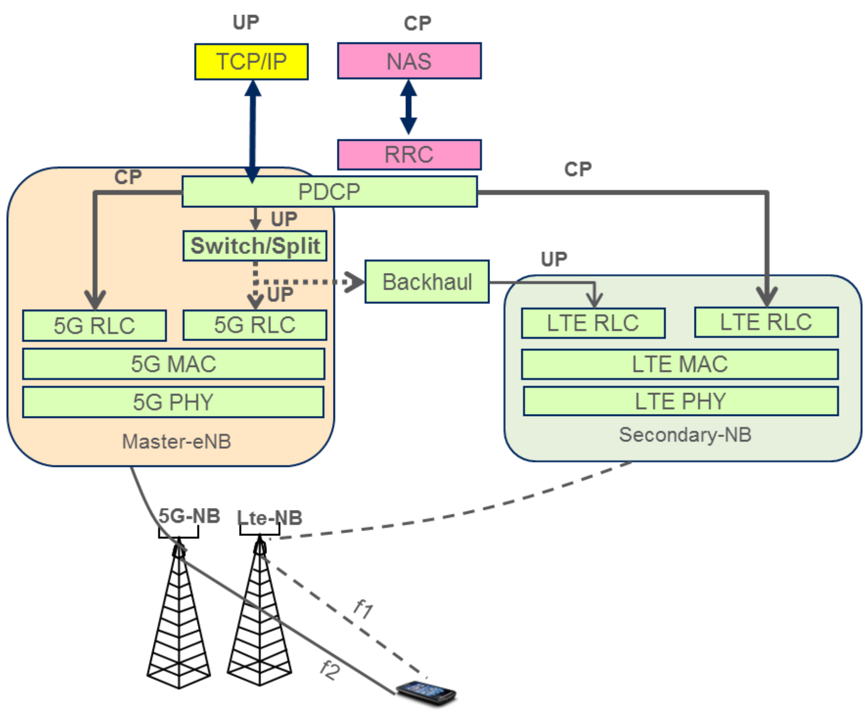


Figure 6-4. Tight interworking options between LTE-A evolution and novel 5G AIVs: Fast UP switching and dual connectivity (downlink example) [MII16-D51].

6.2.2 Resource Management for Network Slices

METIS-II intends to develop a RAN design that fulfils NGMN's vision for the overall native SDN / NFV-based 5G architecture (Software Defined Networking / Network Function Virtualization), as described in the NGMN 5G White Paper [NGM15-WP], which is based on the idea of decoupling the software from the hardware platform of the network as well as to decouple CP and UP. The architecture vision includes a new concept which is described as E2E network slicing. Based on the idea, it will be possible to run multiple logical networks as virtually independent business operations on a common physical infrastructure. Thus, the RAN needs to support differently flavored types of virtual networks (i.e. slice instances) on the same hardware platform with each virtual network optimized for UCs with potentially contradicting KPIs. 3GPP has stated that by now it has to be explored if and which new functionality in the RAN part is needed to support the slicing concept [3GPP15-22891]. With respect to the resource abstraction framework the slicing concept plays an important role. The responsible entity needs enough information about the currently instantiated slices from the CN side to allocate available resources. It can be a new centralized logical element (e.g. an access agnostic Slice Enabler as depicted in Figure 6-5). Alternatively, it is possible to incorporate the features of RM for network slices into the AN-O introduced in the previous section. The allocation of resources is supposed to happen in a way that the QoS requirements of the different services as well as the defined service level agreement (SLA) for a specific slice are considered, as illustrated in Figure 6-5. Briefly, RM for network slices is a novel RM strategy, which enables the sharing of a common RAN (consisting of multiple AIVs) by multiple network slices. This includes an abstraction of RAN resources to perform inter-slice RM with coordination of resource usage by different AIVs and offering a single control point for mobile network. For more information, interested readers are referred to [MII16-D51].

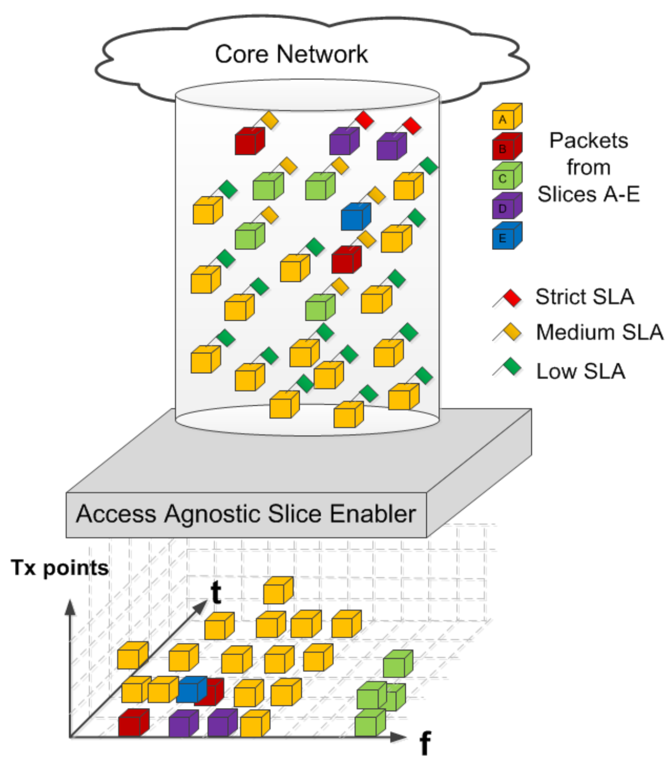


Figure 6-5. Sharing of a common RAN by multiple network slices.

6.2.3 RAN Enablers for Interference Management

Providing an agile framework to handle interference management is a key requirement towards meeting the high capacity and coverage targets for 5G systems. Factors like the expected high density of access nodes reusing the same spectrum, the dynamicity of the topology, the diverse sources of interference from heterogeneous access technologies, and the consideration of multiple 5G services with different KPIs may have a strong influence on both the way that interference management is handled, and how this impacts the RAN protocol design.

Beyond looking into holistic interference management schemes covering dynamic TDD, D2D etc., METIS-II is research three special that deal with interference [MII16-D51]: i) Tunable interference coordination and cooperation schemes (eICIC, coordinated scheduling, joint transmission, and dynamic nomadic node selection) to boost spectral efficiency in a user-centric manner, ii) a hybrid FQAM applicable along the resources of frequency, space, and time dimensions to alter the distribution of inter-cell interference to non-Gaussian and improve cell-edge throughput, and iii) an interference orthogonalization technique by means of a precoding coordination pattern and a spreading factor. All the above mentioned interference management techniques require some level of coordination among base stations that needs to be supported via appropriate signaling and protocols. The following implications of the 5G interference management can be expected in the RAN protocol design:

- New functionalities and interfaces will be required for inter-cell RM and coordination of a dynamic radio topology that is constantly changing due to, e.g., the existence of nomadic nodes. In this scenario, new information elements through X2* interfaces can be needed that provide, e.g., availability of nomadic nodes in a target service region and the backhaul link measurements of nomadic nodes toward possible donor base stations. Collectively, new protocols can be introduced that efficiently handle the activation / deactivation of the nomadic nodes (e.g., in terms of the required signaling overhead) and cope with their possibly changed locations. Besides, a centralized RAN controller may determine the active nomadic nodes based on the aforementioned information elements.
- Backhaul link measurements and activation commands among base stations imply that new signaling elements are needed on the wired or wireless backhaul link using X2* interfaces to support the exchange of information among base stations (e.g., FQAM frequency pool, beams, or subframes; or common scrambling and spreading patterns in case of interference coordination/cancelation inside the base station clusters). Along with the setup of a base station cluster and its operation and management, procedures to agree on a common coordinator for the cluster can be applied.
- In addition to the messages exchanged via the X2* interface, necessary notifications between adjacent cells can be also facilitated via multi-connectivity (e.g., through a lower-frequency AIV that can be used for inter-cell control information). In particular, users being served by small cells can use an additional link with the macrocell to provide interference-related information so that informed decisions on interference management

schemes can be made at the macrocell level. For example, as a result of this control information can facilitate setting the size or location of reserved bandwidth pool for a set of small cells, the amount and identifiers of selected FQAM beams, or the FQAM-based subframes, by sending a broadcast (or customized) message(s) to the cells involved.

- The UEs at the cell border may need to be notified of certain information agreed by a group of cells (e.g., the common scrambling pattern used in the cell cluster or the spreading patterns). For this, a DL control channel like LTE PDCCH could be used.

6.2.4 Novel UE Context Management in 5G

The UE context should be designed to enable the new functionalities required by future 5G devices, such as the support of multiple AIVs and / or numerologies with possibly separate measurements for each AIV / numerology, more frequent inter-AIV switches, the support of high frequencies and various mobility requirements in different UCs. Consequently, the UE context should be adopted to assist in reporting with a better accuracy of the existing parameters such as location, and even reporting new information such as inter-AIV interference. Moreover, the measurement configurations should be adopted to support mobility-based configurations (e.g., a mobility-based measurement interval) and possibly new configurations (e.g., frequency / time / space configurations). Furthermore, the network may need to maintain multiple measurement configurations, one for each AIV or numerology. The indicated functional extensions and changes on the UE context are summarized in Figure 6-6.

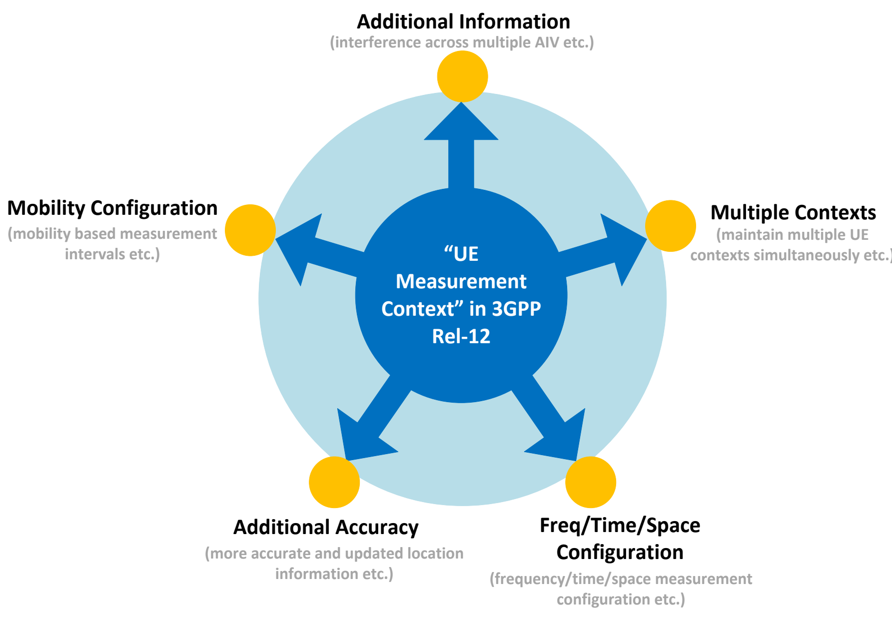


Figure 6-6. Possible changes to the UE measurement context compared to 3GPP Release 12 [MII16-D51].

However, new measurements could exert a negative impact on the UE performance in terms of data gathering, signaling, processing and storage. It is thus important to simplify the UE measurement context as much as possible, for example, a harmonized UE measurement context applicable for different AIVs/numerologies, and take into consideration of, for example, the battery life and utilized memory when the UE measurement context is being designed.

6.3 Functions for Initial Access and Mobility

The CP functions related to initial access and mobility must be able to fulfill a wide variety of requirements such as futureproof design, high energy efficiency, beamforming mobility, higher reliability and tighter integration with legacy AIVs. The following sections describes how some of the 5G CP functions need to be designed in order to handle these requirements. More details can be found in [MII16-D61].

6.3.1 RACH Service Prioritization

In LTE, during initial access, the UE randomly selects one of the random access preambles (out of 64 preambles) informed via broadcasted system information. The current design may create problems e.g. for mMTC applications, where a large number of devices may simultaneously attempt to access the system since the initial access request collisions will lead to additional access delays which may impact services differently.

One solution for the devices with strict latency requirements could be to reserve a set of dedicated preambles for the use of devices with high priority. This solution, however, is not efficient, since the number of RACH preambles is very small (i.e., 64 preambles) which has to be used both for random access and for handover purposes. In order to provide an efficient prioritization mechanism for delay-sensitive services (not relying on the assignment of dedicated preambles) METIS-II is investigating random access solutions to provide some level of access differentiation per service, taking their accessibility requirements into account.

In the considered solution, random access requests associated with delay sensitive services could be configured to apply a combination of preamble signatures at a given random access time slot. The aforementioned approach would enable requests with more strict delay requirements to have higher priority, since combinations of preambles can always be identified by the receiver. This way, requests with higher priority are immune from collisions and the retransmissions, as shown in Figure 6-7.

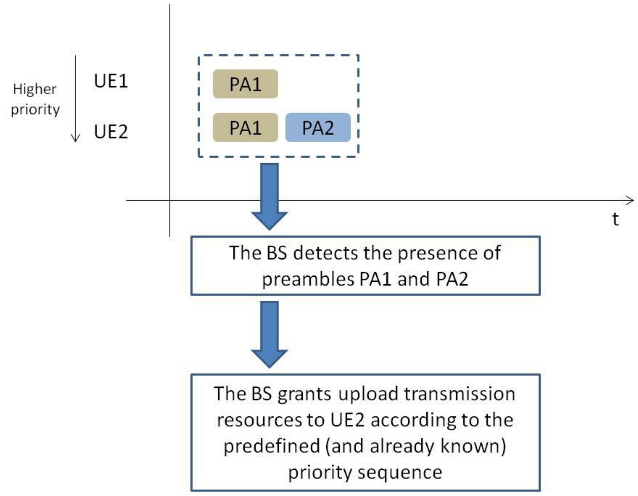


Figure 6-7: Preamble combination for prioritized UE. Preamble is denoted by PA in this figure.

As it is shown in Figure 6-7, the prioritized UE uses a combination of the preamble signatures at one random access time slot to "overwrite" the other preambles. More specifically, UE2 is high priority compared to UE1, thus it sends a combination of the preambles. Preamble PA1 (2 times) and PA2 can be well detected at the RAN receiver respectively. Hence, the receiver detects the preambles PA1 and PA2 and identifies this combination as a high priority request from UE2. Thus, the proposed solution guarantees the priority of a particular request in the random access procedure. The high priority request doesn't need to enter the back-off and retransmission procedure in case of collision, so that the delay caused by collision is minimized for the high-priority request. For more details, please refer to [MII16-D61].

6.3.2 RRC State Management

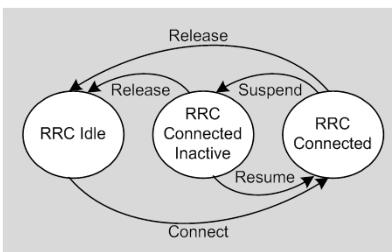
With the trends towards the IoT, it is expected that in 5G there will be even more battery-powered UEs (e.g. sensors, baggage tags, etc.). Therefore, battery efficiency and duration will be essential, especially for devices with limited accessibility (e.g. remote locations, restricted areas etc.), as well as cost. At the same time, the requirement for fast first packet transmission (UL or DL) in mission critical communications may be even more stringent than in current systems. This trade-off between device power efficiency and fast accessibility is often called the "UE sleeping problem". Yet another challenge related to IoT is the large number of devices which infrequently transmit small amounts of data creating high control plane protocol overhead.

A novel RRC state model is proposed to address this problem, relying on a novel state called "RRC Connected Inactive" in addition to "RRC Connected" and "RRC Idle". This novel state explores the principle of "not discarding previously exchanged information" for inactive UEs, meaning that UEs in the new state still keep parts of the RAN context, for instance the security context, UE capability information, etc. In addition to this, signaling is reduced by allowing the

UE to move around within a pre-configured area without notifying the network. During the "RRC Connected Inactive" state, the UE is always having the CN/RAN connection with S1* alive and can be reached by the RAN via notification similar to paging.

The mobility signaling to the CN is avoided when the UE moves within a configured tracking area. The UEs with a low mobility profile can be tracked at cell-level location, thus minimizing the paging load on the AI. Some low latency use cases might require switching the S1* connection to the optimal node immediately when the UE reselects to a new cell or coverage area which is not related to the node currently terminating the S1* connection. This may introduce additional load in X2* interface due to UE Context forwarding.

The new state is also envisioned to be highly configurable with a wide range of discontinuous reception (DRX) cycles (from milliseconds to hours), UE centric location tracking and service-tailored optimizations related to the transition to "RRC Connected". Figure 6-8 shows the three considered states and related state transitions.



	4G (Transition from idle to connected)	5G (Transition from connected inactive to connected)	Gain
Signaling overhead	13x RRC messages* 8x S1AP messages** *7 if the CN has kept some context information **3 if the CN has kept some context information	2x RRC messages Network signaling related to context fetching* * Though this may be negligible if UE returns to same area	At least 70% signaling reduction
Control Plane Latency	RA delay + 3.5x radio RTT + S1 RTT	RA delay + radio RTT	Reduction by 2x radio RTT + S1 RTT

S1AP: S1 Application Protocol, RTT: Round trip time

Figure 6-8. Proposed novel RRC state model in 5G, and its benefits. Roundtrip time is denoted as RTT in this figure.

In the novel model, transitions from "RRC Idle" to "RRC Connected" are expected to occur mainly when a UE for the first time attaches to the network, or as a fallback case when the devices and / or the network cannot use the previously stored RAN context. Transitions from "RRC Connected Inactive" to "RRC Connected" are expected to occur often and are hence optimized to be fast and lightweight in terms of signaling. This is achieved by keeping the CN /

RAN connection alive during inactivity periods and reducing the amount of RRC signaling needed to resume an existing inactive connection via the usage of a RAN context ID – an approach that is inspired by the suspend / resume procedure to be defined for idle state UEs in LTE in Rel. 13 [3GPP15-151621]. More details can be found in [SMS+16] and [MII16-D61].

6.3.3 RAN-based Paging

Due to the expected massive number of devices and denser deployments in 5G, paging may significantly increase the load on both the AI and the CN / RAN interface, thus requiring new solutions for efficient paging and UE location tracking. One option could be the introduction of such functionalities in the RAN, e.g. in the form of a hierarchical location tracking where the CN tracks the registration of UEs in "RRC Connected Inactive" only on the level of groups of RAN locations, whereas the RAN tracks these on cell-level granularity. This could involve a lightweight signaling procedure terminated in the RAN, using a security handling mechanism based on retaining and updating the security context from the last attach procedure. The considered hierarchical location tracking approach would imply moving part of the paging and mobility related functions from CN to RAN, hence changing the CN / RAN split as compared to EPS. The expected benefits from RAN-based paging are summarized in Figure 6-9.

	LTE-A (Paging an RRC Idle UE from CN)	**5G** (Paging a UE in RRC Connected Inactive from RAN)	**Gain**
Signaling overhead (for the transfer of the first packet)	State transition: 13x RRC, 8xS1AP, Paging (MME-S-GW): 2 x S11, Paging Core: S1AP x 4G-cells per TA-list, Paging RRC: RRC x 4G-cells per TA-list, Paging Response: 1 x RRC, 1 x S1AP	State transition: 2 x RRC messages Paging: 2 x RRC messsages Network signaling related to context fetching	> 85% signaling reduction The gain increases with the number of cells per tracking area
User Plane Latency (from packet arrival at S-GW until packet reaches UE)	RA delay + 4.5x radio RTT + 2xS1 RTT + S11 RTT + packet tx delay	RA delay + 1.5x radio RTT + packet tx delay	Reduction by 2.5x radio RTT + 2xS1 RTT + S11 RTT

S1AP: S1 Application Protocol, RTT: Round trip time

Figure 6-9. Expected benefits from RAN-based paging.

Although the RAN based paging seems to minimize CN involvement in the paging procedure, it is done at the cost of added RAN complexity and a higher signaling load on the X2* interface. RAN based paging introduces the need for forwarding the notification of incoming data as well as forwarding the actual UP data from the last serving cell to the current serving cell. By limiting the RAN tracking area to a single UE-centric node would simplify the inter-node interfaces but would also limit the area within which UE-based cell reselections are allowed. Further details can be found in [MII16-D61].

6.3.4 Mobility Management

In LTE, the mobility procedure in RRC Idle state is UE-based and optimized for power saving and minimized signaling while in RRC Connected state mobility is network based and optimized for service continuity. METIS-II believes the same approach should be applied to 5G as well, i.e. mobility procedures should be individually optimized for RRC Connected and RRC Connected Inactive state, handling active data transmission and low activity periods, respectively. Mobility procedures for RRC Idle state may be needed mainly for fault management due to radio link failures and in some fallback procedures where the de-registered UE needs to perform PLMN selection and re-attach to the network.

How to perform the handover procedure to handle to very strict requirements is of major importance in 5G. For latency-critical use cases it is important to minimize the service interruption due to mobility events for both ideal and non-ideal backhaul scenarios. If the requirement of interruption time is down to 0 ms, then make-before-break mobility events are required implying handovers using dual/multi-connectivity. However, the interruption time requirement in 5G might not be the same for all use cases. Therefore, we propose both break-before-make and make-before-break handover procedures to be considered in 5G. The break-before-make procedure might be a natural consequence of single connectivity, whereas, the make-before-break handover is a natural consequence and use case of multi-connectivity.

The beam forming mobility design should support a fast switching / tracking of the communication beam to combat rapid changes in link quality. The design should be able to exploit the availability of multiple overlapping beams used for the communication with a single UE. Further, the beam management should have a minimum impact to the RRC layer. One solution to fulfill these requirements is the idea of cluster-set based mobility, which is a set of nodes that the UE can detect and which are prepared in advance for a fast re-routing of the signaling and user data. More details on this can be found in [MII16-D61].

Mobility with the use of massive beamforming causes new challenges for 5G at higher carrier frequencies. Due to the lean design of the DL reference symbols, the reported mobility measurements of the DL may carry limited measurement information for network handover procedures. One approach to overcome the limitation is to use the UL measurements for mobility. For the UL measurements, METIS-II proposes a new scheme where each UE (at least

UEs in RRC Connected) directionally broadcasts a sounding reference signal (SRS) in a time-varying direction that continuously sweeps the angular space. Each potential serving cell scans all its angular directions and monitors the strength of the received SRS along with its variance to capture better the dynamics of the channel. In order to take advantage of the UL measurements for mobility, the neighboring cells needs to exchange the UL measurement information, thus increasing the signaling load on the X2* interface.

In LTE, the tracking area management needed to locate UEs in RRC Idle state is centralized to the Mobility Management Entity. To allow for always-connected UEs in 5G, the location tracking could be distributed and done in the RAN, since the mobility anchor and S1* interface connection to CN is available also during the low activity periods. In this case, the UEs in RRC Connected Inactive state do not create location update signaling towards a CN when reselecting a new cell during a low activity period.

To minimize the need for S1* path switching and UE context transfers in the RAN due to cell reselections during RRC Connected Inactive state, the S1* path(s) can remain terminated in the last serving node where the UE was last time in RRC Connected state. The last serving node takes the role of mobility anchor, which allows keeping the CP and the UP connections unmodified and active towards the CN. The UE performs cell reselection and may inform the network about the new cell identity, but the network may decide not to perform the last serving node change. Instead, the last serving node can establish a bi-directional X2* tunnel which is used for transferring the UP data from the mobility anchor to the serving node. Some low latency use cases may require switching the S1* connection to the optimal node immediately when the UE reselects to a new cell or coverage area which is related to the node currently terminating the S1* connection. Therefore, for some URLLC 5G use cases, the mobility during Connected Inactive state may cause frequent S1* interface path switching. Further discussion can be found in [MII16-D61].

6.4 Summary

The previous sections provided a brief summary on key functional design considerations from METIS-II, which are being developed and evaluated in WP4 [MII16-D41], WP5 [MII16-D51], and WP6 [MII16-D61]. Table 6-2 summarizes these functional design considerations for 5G, and highlights their key benefits, the differences to LTE-A evolution, and the main implications on the overall 5G RAN design.

The key functional considerations move towards several directions and benefit the network in various ways. In brief, the benefits include better coverage and capacity (i.e., beam-centric design, relaying and self-backhauling, interference management, multi-AIV interworking), increased energy efficiency (i.e., lean design, energy efficient RAN moderation, optimized UE context measurement), and increased flexibility (i.e., AIV configuration, traffic steering, slicing, etc.). Additionally, signaling overhead is reduced using efficient and optimized mobility management, a new RRC state model, and optimized initial access. Apart from the reduction of the signaling overhead, the novel functionalities also increase the reliability and reduce the latency.

Major differences compared to LTE-A include service-oriented designs (e.g., targeted enumerations of the AIVs for certain services, optimized random access allowing for service prioritization, a service-tailored RRC state transition handling), all of which provide the basis of network slicing in 5G. Additionally, the novel network design, contrary to that of LTE-A, may enable the system information distribution and the reference signals transmission only when needed. Furthermore, compared to LTE-A, optimized RM techniques are being incorporated for providing efficient traffic steering and interference management, whereas certain functionalities such as D2D and self-backhauling, which in previous deployments have been added on in mature phases of the LTE-A, are natively integrated in the 5G system. Finally, the UE measurements, and mobility management will in 5G focus on the new needs with multiple AIVs available, and an extensive use of beamforming.

Table 6-2. Summary on key functional design considerations for 5G.

5G Functional Design Paradigm	Key benefits	Key difference to LTE-A evolution	Implication on overall RAN design
Beam-centric Design	Better coverage, capacity and data rates in higher bands	Narrow beams possibly swept instead of omni-directional cells	Major; all control signals beamformed; all mobility and initial access procedures need native beam-centric design
Lean and Future-proof Design	Energy efficiency and future-proofness, potentially also improved C-plane scalability	Reference signals not always on, not full band, not all subframes	Significantly more configurable reference signals and mobility proc.
Energy Efficient Design and RAN Moderation in 5G	Reduction in overall network energy consumption with good throughput trade-off	Exploitation of coordination schemes to attain high energy efficiency both in FDD and dynamic TDD systems leveraging on the improved power consumption models of 5G nodes	New information elements over the X2* interface to indicate the load and interference information on a TTI-level; new information elements over X2* to indicate the type and level of cooperation; a centralized entity for coordinated scheduling

Native Relaying, Self-backhauling and D2D support in 5G		Efficient support of 5G services that can benefit, e.g., from capacity, resource reuse, power consumption and coverage gains offered by these technologies	Native integration since the beginning of 5G system design (e.g., in terms of CP functionalities, frame structures, etc.) rather than an add-on feature on top of an already mature system like LTE	CP and UP functionalities ranging from PHY to higher layers should consider native D2D and self-backhauling support
Multi-AIV Resource Mapping	Dynamic Traffic Steering	Holistic and agile RM; higher overall reliability;	Traffic steering performed on comparatively at lower protocol stack layer; dynamic traffic steering instead of handover	New control information elements and steering options between protocol layers needed, poss. impact on Uu
	AIV abstraction and AIV-agnostic functionality	Efficient and lean RM with multiple AIVs and dynamic topologies	Multiple AIVs and dynamic topologies in 5G require an efficient way to deal with them from RM perspective	Split between AIV-specific vs. AIV-agnostic functionalities; message exchange over X2* for edgeless user experience
	Flexible frame structure with variable TTI size	Simultaneous support of users with very different service requirements	TTI size tailored per user and per scheduling instance according to data rate, latency, reliability and coverage needs	PHY support of flexible frame structure with variable TTI size
RM for Network Slicing		Possibility to share a common RAN for multiple businesses and services with diverging requirements	Network Slicing is a new feature which is not part of LTE-A	New RM concepts required to implement slice aware resource assignment; possible new entity that performs slice-overarching RM
Multi-AIV Interworking on Fast Time Scale		Tight interworking increases user bit-rate and connection reliability	Avoidance of inter-RAT hard handover which causes a transmission interruption; also fast PDCP level aggregation/switch between AIVs, i.e. LTE-	Fast addition and deletion of a new CP connection in dual connectivity to a UE along with lightweight signaling to support ultra-reliability; new signaling for AIV quality metric; metrics for enabling both load balancing and

		A and 5G	traffic steering between LTE and 5G beneficial, but require new meas. over X2*
RAN Enablers for Interference Management	Higher cell-edge use throughput, larger capacity and better coverage	Advanced cooperative interference mgmt. techniques targeted at dynamic topologies and dense deployments, for instance with flexible UL/DL TDD	RAN impact is mostly characterized by the need for signaling and procedures over the wired or wireless backhaul using X2* interface to support the exchange of information among cooperating BSs
Novel UE Measurement Context in 5G	Reduced overhead, enhanced energy efficiency	Functional extensions and changes in the UE measurement context	New information and configurations in the UE measurement context; Option that a UE may maintain multiple measurement contexts
A Novel RRC State Model	Reduced UE power dissemination, C-plane latency and CN/RAN signalling, esp. suitable for bursty connectivity and massive access	UEs are always connected from a CN perspective; significantly larger possibilities for service-spec. configuration	Context fetching needs to be specified and supported. Novel mobility procedures for new state to be defined
Service Prioritization at Initial Access	Service differentiation already at first access; lower latency for mission-critical services	Different levels of service prioritization for diverse sets of delay requirements without reserving resources for certain service classes	New MAC procedures required for RACH to enable service prioritization; signalling to higher layers
Mobility Management	Mobility with very low interruption delays and efficient beam-forming mobility	Support for extreme low interruption handover and functions to handle massive beam-forming	Major; beamforming mobility requires new set of measurements and signalling; new mobility procedures to handle low interruption delay HO
RAN-based Paging	Reduced CN/RAN signalling, reduced C-plane latency	In LTE paging is a CN function, which is now moved into the RAN	Entire re-design of paging functionality, signalling etc., change of usage of CN/RAN interface

7 Key RAN Design Questions Addressed

In the METIS-II project proposal and description of work, the partners had posed 11 key 5G RAN design questions seen to be most important to be addressed during the course of the project. Table 7-1 below now provides a short summary on the status of the work w.r.t. these key questions, and the information where the reader can find details on the work. In case certain key RAN design questions could not yet be fully answered, it is briefly listed when these will be tackled, and in which future report or deliverable a final answer will be provided. It is worth noting that, as shown in Figure 1-3, various deliverables in the first year of the project cover the draft considerations (i.e., D2.2, D4.1, D5.1, and D6.1), while the final considerations will be provided towards the end of the project (i.e., D2.4, D4.2, D5.2, and D6.2).

Table 7-1. Status of the METIS-II work on answering key 5G RAN design questions.

No	Key RAN Design Aspect / Question
1	**What is the general spectrum usage foreseen for 5G?** A brief summary on the general spectrum usage foreseen for 5G is given in Section 4.1.1. More details can be found in [MII16-D31].
2	**Given the various characteristics of different spectrum bands, which band should be used for what type of service, air interface and how much spectrum needs to be made available for mobile communications in the different bands?** A brief summary on which band should be used for what type of service is given in Section 4.1.1, with more details in [MII16-D31]. Initial considerations on spectrum needs in different bands are given in [MII16-R31]. A more detailed spectrum demand analysis will be provided in deliverable D3.2, to be published June 2017. The question of which bands should be used for which service was additionally covered in [MII15-R41, MII16-D41], by determining which overall set(s) of AIVs, e.g. operating in different spectrum bands and / or tailored towards certain services, will be most suitable to address the overall 5G requirements space.
3	**Which air interface variants are expected to be introduced in the context of 5G, and which are to be evolved from existing standards?** A preliminary answer to this question is given in [MII15-R41], where it is explained why legacy technology and its likely evolution will not be able to meet many of the 5G requirements. The report further describes specific physical layer components and selected AIVs building upon these components, meeting some or many of the 5G requirements and contributing to the METIS-II overall AI design goals. This work is further extended in [MII16-D41], where overall AI proposals being considered for 5G are described in detail, as well as their relationship with existing standards. A further

	assessment and down-selection of suitable AIVs and possible overall AI frameworks is ongoing and will be documented in D4.2, to be published in April 2017.
4	**How many different novel and legacy air interface variants should different devices support? Which forms of concurrent connectivity (e.g. multi-standard and multi-cell connectivity, concurrent device-to-device and device-to-infrastructure connectivity) will be required in 5G?** A commonality analysis of different METIS-II AI proposals was carried out in [MII16-D41]. Following this analysis, an initial assessment was performed of the AI proposals (grouped around subjacent waveform technologies) using the harmonization KPIs conceived within METIS-II, where various forms of AI aggregation were identified and their features discussed. This has laid the foundation for understanding the impact on device complexity and whether a device should be able to serve multiple AIVs simultaneously, as well as whether a device should be able to conduct a transmission to the infrastructure and to another device simultaneously, again in the context of user plane design. These investigations will be continued towards D4.2, to be published in April 2017. Further, the stated questions have been touched from the perspective of which form of multi-connectivity or concurrent device-to-device and device-infrastructure connectivity is actually beneficial in the context of holistic RM. Furthermore, possible implications on the device complexity have been taken into account for the investigation on potential UE context extensions, as described in Section 6.2.4. Further details on draft considerations in this direction are captured in [MII16-D51].
5	**How tightly are novel air interface variants expected to be integrated with each other and with legacy technologies (e.g. LTE evolution and Wi-Fi), to which extent should they be harmonized or have common functionality in the protocol stack, and on which level should different transmission forms be aggregated?** The METIS-II partners have concluded that the integration among LTE-A evolution and novel AIVs, or the integration among multiple novel AIVs, should be possible on RAN level [MII16-D61]. In this respect, specific envisioned forms of UP aggregation and CP integration have been described in detail in Section 5.4. Regarding harmonization, the general view is that among novel AIVs, a large extent of protocol stack function harmonization should be strived for (i.e. at least a harmonized MAC and higher layers) [MII16-D41]. Among LTE-A evolution and novel AIVs, harmonization has to be carefully traded against possible backwards-compatibility constraints imposed towards 5G technology. Here, it is assumed that at least the PDCP layer and above could follow a common evolution, as stated in Section 5.4. These intermediate considerations are expected to be concluded in D4.2, D5.2, D6.2 and D2.4 throughout April – June 2017.
6	**How can one efficiently handle interference in an ultra dense environment? What kind of information is required, at what time scale and how fast the system must react?** This question is still being investigated. Various TeCs constituting the Agile RM Framework of METIS-II are targeting interference management as first exemplified in [MII15-R51]. It is emphasized here that the way of handling interference depends on the operational scenario and use case. For example, when a centralized RAN approach can be implemented, it is possible to design a centralized interference management system exploiting detailed channel state information (CSI) and operating on a very short (e.g., 1

	ms) timeframe. The CSI may be fed back from the users to one or more access points (which might be related to different AIVs) dynamically. The applicability of the available information and the related timescales for information exchange and network configuration updates are being investigated and have been documented in [MII16-D51]. Furthermore, the concept of lean design for common signals reduces the amount of interference, which is an important enabler for the 5G system to handle ultra-dense environments. This has been summarized in Section 6.1.3 and is detailed in [MII16-D61].
7	**What will be considered as "resource" in a 5G system? How can we manage these resources effectively in order to achieve the 5G KPIs?** As captured in [MII16-D51], it is expected that in 5G the notion of a resource may be expanded beyond classical radio resources to also include further dimensions of spectrum resources (e.g. unlicensed bands, spectrum sharing), transmission points along with radio frequency (RF) equipment and their soft capabilities, such as processing resources, storage and memory resources, transport network resources and energy. With respect to how this extended notion of resource will be managed efficiently, preliminary considerations have been presented in [MII16-D51] and summarized here in Section 6.2. A conclusion on this topic is expected to be provided in D5.2 in May 2017.
8	**On which time scale should certain 5G radio access network functionality (e.g. radio RM, radio resource control, mobility) operate, and consequently, how should the necessary functionalities be best abstracted, grouped and tackled in standardization and implementation?** The general trend is that, in 5G, many functionalities are expected to be handled on a faster time scale than in legacy systems. For instance, it is envisioned in METIS-II to enable mobility and multi-connectivity among LTE-A evolution and novel AIVs on RAN level, inherently allowing for a faster setup of new multi-connectivity constellations and switching among these, see Section 5.4.2. Further, it is envisioned that traffic steering among different AI technologies, which was so far done via hard handover, is performed on lower protocol stack layers and consequently on a much faster time scale, as detailed in Section 6.2.1. Additionally, taking into account the overall AI consisting of multiple AIVs, the functionalities, e.g., related to RM, may be grouped in terms of intra-AIV schemes and AIV-overarching schemes, see Section 6.2 and [MII16-D51]. The most suitable time scale of procedures related to RRC state handling, mobility and system access is investigated and documented in [MII16-D61]. Note that the aforementioned aspects are initial considerations, and will be concluded and further detailed in D5.2, D6.2 and D2.4, to appear in the time frame April – June 2017.
9	**How will the concepts from dynamic spectrum management interwork with the control plane architecture (new network elements and interfaces for this purpose and/or some level of integration to the control plane design)?** The dynamic spectrum management is to provide the information on which parts of the spectrum are currently accessible for the 5G system also related to the location and available infrastructure. Also, as it is not expected that all radio nodes are equipped by same RF parts and antenna types, access node and device capabilities are also relevant. The "Holistic Spectrum Management Architecture", which will be part of METIS-II deliverable D3.2 to be available in June 2017, is going to consider these aspects.

10	**What will be the network elements and interfaces in the 5G system architecture and, assuming these, how would these interfaces look like, i.e. which functionalities will they have, which programmability level will be adopted, what level of openness, what level of abstraction, etc.?** METIS-II has already started to investigate the requirements for the overall 5G logical architecture and, more specifically to the RAN control plane. An important assumption taken in METIS-II is the logical split between the RAN and CN (and Service Layer) functions, see Section 5.2. In this respect, the project has designed a new connected state optimized for inactivity periods between small packet transmissions, and RAN-based paging solutions to address densified deployments, both approaches implying a shift of functionality from CN to RAN, see Section 6.3 and [MII16-D61]. Regarding intra-RAN interfaces, it is assumed that an evolved X2* interface between access nodes will be required, for various reasons detailed in Section 5.3. It is expected that this interface will also be crucial for agile interference management in 5G, as listed in Section 6.2.3 and described in further detail in [MII16-D51]. Regarding possible function splits in the RAN for different physical deployments, various options have been discussed in Section 5.5. This topic is expected to be concluded in D2.4 in June 2017. The preferred level of programmability is still under investigation. For the RAN, it may be assumed as a baseline (for comparison purposes) the current programmability levels enabled by Operations, Administration and Maintenance (OAM) interfaces where centralized (OSS-based) SON is one example of how to program and control RAN features. Nevertheless, it is anticipated that extensions to such capabilities will be investigated to achieve a more flexible and adaptable control plane functionality. For that purpose, programmability requirements will be proposed and solutions derived in D6.2, to appear in June 2017.
11	**What type of control and user plane functionalities should be centralized or distributed depending on the 5G use cases associated to them? Out of these functionalities, what are the most promising candidates to be implemented as virtual network functions?** Preliminary considerations on the centralization and distribution of network functions have been provided in Section 5.5, for instance proposing specific function splits between physical network entities for different deployment scenarios. A use-case / physical architecture specific proposal of function splits will be provided in D2.4 in June 2017. The degree of centralization and the associated benefits also constitute an important aspect for the RM functionalities within the agile RM framework, which considers both centralized and distributed control functions. The work within the agile RM framework may also impact the mechanics of CP / UP split. It is, as well, seen important to enable RM for network slices, which adds an additional dimension in terms of business requirements. In this direction, initial considerations have been provided in Sections 6.2.1 and 6.2.2, with further details in [MII16-D51].

8 Summary and Outlook

This deliverable has captured the draft 5G RAN design considerations at the mid-point of the METIS-II project. It has provided an insight into the current METIS-II view on the 5G air interface (AI), which is expected to be composed of multiple AI variants (AIVs), including both evolved legacy technology such as LTE-A, as well as novel AIVs, for instance tailored to different services and bands. The document has furthermore summarized the key consensus found so far on the overall 5G architecture, for instance the view on a logical CN / RAN split, on the network interfaces allowing to integrate new AIVs with evolved LTE-A, the likely mapping of logical to physical architecture for different deployments, and architectural enablers for network slicing. In a dedicated chapter, the reader has further obtained an overview on various functional design considerations of the METIS-II project, with a brief summary of the key expected benefits and implications on the overall RAN design.

As could be seen from Chapter 7, the METIS-II partners have already obtained a good common understanding of the 5G RAN design, and have been able to provide preliminary answers on most of the key RAN design questions that were posed at the beginning of the project. In the remainder of the project, it is now important to provide more technical details to the developed solutions, and in particular provide numerical evaluations to be able to quantify the benefits of the considered schemes when put in conjunction, and to see whether the developed concepts fulfill the overall 5G needs as stated in [MII16-D11]. It further has to be cross-checked that the METIS-II solutions are capable to scale to any requirement extremes, adapt to any of the considered use cases, and work with any of the physical deployment architectures defined in Section 5.5. All these aspects will be captured in final deliverables towards the end of the project, according to the timeline depicted in Section 1.1.

References

[3GPP13-36842] 3GPP TR 36.842 “Study on Small Cell enhancements for E-UTRA and E-UTRAN” Release 12, v12.0.0 (2013-12).

[3GPP15-153651] 3GPP TR S2-153651, “Proposal for Study on a Next Generation System Architecture”, 2015.

[3GPP15-151621] 3GPP, RP-151621, “New Work Item: NarrowBand IOT (NB-IOT)”, September 2015

[3GPP15-23720] 3GPP TR 23.720, Technical Specification Group Services and Systems Aspects, “Architecture enhancements for Cellular Internet of Things (Release 13)”, Nov. 2015.

[3GPP15-22891] 3GPP TR 22.891, “Study on New Services and Markets Technology Enablers”, Release 14, v1.0.0 (2015-09), September 2015.

[3GPP15-36300] 3GPP TS 36.300, “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2”, December 2015.

[3GPP15-36889] 3GPP TR 36.889, “Feasibility Study on Licensed-Assisted Access to Unlicensed Spectrum”, December 2015.

[3GPP16-22011] 3GPP TS 22.011 “Technical Specification Group Services and System Aspects;

Service accessibility (Release 14), March 2016

[3GPP16-22864] 3GPP TR 22.864, “FS_SMARTER - Network Operation”, March 2016.

[3GPP16-38913] 3GPP TR 38.913, Study on Scenarios and Requirements for Next Generation Access Technologies (Release 14), v0.2.0 Feb. 2016.

[3GPP16-160043] 3GPP RP-160043, Summary of “[5G-AH-06] Fronthauling,” Telecom Italia, et al.

[5GARCH16-WP] 5G PPP, Architecture White Paper, May 2016.

[5GEx16-WP] 5GEx, White Paper “5GEx Multi-domain Service Creation - from 90 days to 90 minutes”, March 2016.

[5GNOW] 5GNOW project, www.5gnow.eu

[5GPPP16-UCM] M. Maternia and S. El-Ayoubi (editors), “5G PPP Use Cases and Performance Evaluation Models”, see https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-use-cases-and-performance-evaluation-modeling_v1.0.pdf, April 2016.

[5GPPP16-WP] 5G PPP, White Paper, “5G empowering vertical industries”, March 2016.

[5GREEN] 5GREEN project, see https://www.researchgate.net/publication/260126161_5GrEEn_Towards_Green_5G_mobile_networks

[5GN15-D31] 5G NORMA, Deliverable D3.1, "Functional Network Architecture and Security Requirements", December 2015.

[5GXHVLC] 5G-XHaul, "Data rate and timing requirements on 5G Transport networks" Vaia Kalokidou (University of Bristol), Jens Bartelt (Dresden University of Technology), Albrecht Fehske (Airrays GmbH), Nicola Vucic (Huawei). Presentation on the joint meeting in Valencia.

[CHAR16-D11] 5G PPP CHARISMA, Deliverable D1.1, "intelligent, distributed low-latency security C-RAN/RRH architecture", April 2016.

[EARTH] EARTH project, see https://www.ict-earth.eu/

[ETSI-NFV] ETSI Industrial Specification Group (ISG) Network Function Virtualisation (NFV), http://www.etsi.org/technologies-clusters/technologies/nfv

[ETSI14-MAN] ETSI ISG NFV, GS NFV-MAN 001, "Network Functions Virtualisation (NFV); Management and Orchestration", V1.1.1, December 2014.

[iJOIN-D22] iJOIN D2.2 "Definition of PHY layer approaches that are applicable to RANaaS and a holistic design of backhaul and access network" (Oct 31st, 2014).

[iJOIN-D23] iJOIN D2.3 "Final definition and evaluation of PHY layer approaches for RANaaS and joint backhaul-access layer" (Apr 30th, 2015).

[ITU15] International Telecommunication Union – Radio (ITU-R) Working Party WP 5D: Draft New Recommendation ITU-R M.2083-0 (09/2015), "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond", September 2015.

[METIS] ICT-317669 METIS project, see http://metis2020.com

[MET14-D53] ICT-317669-METIS, Deliverable D5.3 Version 1, "Description of the spectrum needs and usage principles", August 2014.

[MET15-D15] ICT-317669 METIS, Deliverable 1.5, "Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations", April 2015

[MET15-D64] ICT-317669 METIS, Deliverable 6.4 Version 2 "Final report on architecture", January 2015.

[MII15-R11] ICT-671680 METIS-II, Report 1.1 Version 1, "Preliminary refined scenarios and requirements, and consolidated use cases", September 2015.

[MII15-R21] ICT-671680 METIS-II, Report 2.1 Version 1, "RAN design guidelines and methodology", September 2015.

[MII15-R31] ICT-671680 METIS-II, Report 3.1 Version 1, "Preliminary spectrum scenarios and justification for WRC Agenda Item for 5G bands above 6 GHz", October 2015.

[MII15-R41] ICT-671680 METIS-II, Report 4.1 Version 1, "Preliminary air interface analysis and user plane design considerations", November 2015.

[MII15-R51] ICT-671680 METIS-II, Report 5.1 Version 1, "Preliminary synchronous control functions and resource abstraction considerations", November 2015.

[MII16-D11] ICT-671680 METIS-II, Deliverable D1.1 Version 1, "Refined scenarios and requirements, consolidated use cases, and qualitative techno-economic feasibility assessment", January 2016.

[MII16-D31] ICT-671680 METIS-II, Deliverable D3.1 Version 1, "5G spectrum scenarios, requirements and technical aspects for bands above 6 GHz", May 2016.

[MII16-D41] ICT-671680 METIS-II, Deliverable D4.1 Version 1, "Draft air interface harmonization and user plane design", May 2016.

[MII16-D51] ICT-671680 METIS-II, Deliverable D5.1, "Draft Synchronous Control Functions and Resource Abstraction Considerations", May 2016.

[MII16-D61] ICT-671680 METIS-II, Deliverable D6.1, " Draft asynchronous control functions and overall control plane design", June 2016.

[MII16-WP] ICT-671680 METIS-II, White Paper on "Preliminary Views and Initial Considerations on 5G RAN Architecture and Functional Design", March 2016.

[NGM15-CR] NGMN, Deliverable "Further Study on Critical C-RAN Technologies", March 2015.

[NGM15-WP] NGMN, "5G White Paper", March 2015.[ITU14] IMT Vision, "Framework and overall objectives of the future development of IMT for 2020 and beyond", ITU, Feb. 2014.

[RBM+15] P. Rost, I. Berberana, A. Maeder, et al.: Benefits and challenges of virtualization in 5G radio access networks", IEEE Commun. Mag., December 2015.

[SMS+16] I. Da Silva, G. Mildh, M. Säily, S. Hailu, "A Novel State Model for 5G Radio Access Networks", IEEE ICC Workshop, 2016

[WRB+14] D. Wübben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis. "Benefits and impact of cloud computing on 5G signal processing". IEEE Signal Processing Magazine (Volume 31, Issue 6), Nov. 2014.

[WZZ15] J. Wu, Y. Zhang , M. Zukerman, "Energy-Efficient Base-Stations Sleep-Mode Techniques in Green Cellular Networks: A Survey", IEEE Communication Surveys & Tutorials, Vol. 17, No. 2, Second Quarter 2015.

A Appendix

A.1 Data Rate Requirements of different Fronthaul Interface Options

Models for the data rate requirements of the split options M0-M9 in Section 5.5.2 have partly already been discussed in the literature [iJOIN-D22, iJOIN-D23, 5GXHVLC, WRB+15].

The selected models to calculate the data rate requirements for each of the split options are listed in Table A-1, while the parameters are listed in Table A-2. A fully loaded system with a single user is assumed.

Table A-1. Models used within the functional split analysis (downlink only).

Split option	Downlink	
M1	$DR_{DL}^{M.1} =$ $2 * N_A * N_{Q,1} * OF * f_S * \gamma =$ $2 * N_A * N_{Q,1} * OF * (N_{FFT} + N_{CP}) * T_S^{-1} * \gamma =$ $2 * N_A * OF * (N_{FFT} + N_{CP}) * N_{SYMB}^{SUB} * N_{Q,1} * T_{SUB}^{-1} * \gamma$	Scaling with bandwidth and antennas
M2	$DR_{DL}^{M.2} = 2 * N_A * N_{SC} * N_{SYMB}^{SUB} * N_{Q,2} * \gamma * T_{SUB}^{-1}$	
M3	$DR_{DL}^{M.3} = N_L * N_{SC} * N_{SYMB}^{SUB} * Q_m * \mu * \gamma * T_{SUB}^{-1}$	Scaling with user data rates
M4	$DR_{DL}^{M.4} = N_L * N_{SC} * N_{SYMB}^{SUB} * C * \mu * \gamma * T_{SUB}^{-1}$	
M5-M8	$DR_{DL}^{M.5-8} = N_L * N_{SC} * N_{SYMB}^{SUB} * C/(1+r) * \mu * \gamma * T_{SUB}^{-1}$	

Table A-2: Explanation of parameters used in Table A-1.

Parameter	Explanation
$\mathrm{N_A}$	Number of antennas
$\mathrm{N_Q}$	Number of quantization bits for each signal component
OF	Oversampling factor
f_S [MHz]	Sampling frequency at FFT/IFFT output

γ	Transport overhead
N_{FFT}	FFT/IFFT size
N_{CP}	CP size
T_S [ms]	Symbol duration (without CP)
N_L	Number of spatial layers transmitted by a user
N_{SC}	Number of available subcarriers for transmission (scales with system bandwidth B)
N_{SYMB}^{SUB}	Number of OFDM symbols per subframe
µ	Utilization factor
T_{SUB} [ms]	Subframe duration
Q_m	Number of bits carried per symbol (depends on the modulation)
R_C	Coding rate (amount of redundancy added by FEC)
C	Number of data bits per symbol (depends on the modulation scheme and the coding rate) $C = Q_m * R_C$
r	Retransmission rate

The exemplary parameters listed in Table A-3 were used to calculate the data rate requirements displayed in Figure 5-8.

Table A-3: Example parameters used in calculation.

Parameter		Value
$\mathrm{N_A = N_T = N_R = N_L}$		2
$\mathrm{N_Q}$	$\mathrm{N_{Q,1}}$	15
	$\mathrm{N_{Q,2}}$[6]	7
	$\mathrm{N_{Q,3}}$	1
OF		1
f_S [MHz]		30.72
γ		1.33

[6] Proposed in [NGM15-CR] and [iJOIN-D22]

N_{FFT}	2048
N_{CP}	144
T_S [µs]	66.6
N_{SC}	1200
N_{SYMB}^{SUB}	14
µ	1
T_{SUB} [ms]	1
Q_m	6
R_C	0.75
C	4.5
r	0.1