



5G PPP Architecture Working Group

View on 5G Architecture

Version 3.0, February 2020

Date: 2020-02-14

Version: 3.0

DOI 10.5281/zenodo.3265031

URL <http://doi.org/10.5281/zenodo.3265031>

Abstract

The 5G Architecture Working Group as part of the 5G PPP Initiative is looking at capturing novel trends and key technological enablers for the realization of the 5G architecture. It also targets at presenting in a harmonized way the architectural concepts developed in various projects and initiatives (not limited to 5G PPP projects only) so as to provide a consolidated view on the technical directions for the architecture design in the 5G era.

The first version of the white paper was released in July 2016, which captured novel trends and key technological enablers for the realization of the 5G architecture vision along with harmonized architectural concepts from 5G PPP Phase 1 projects and initiatives. Capitalizing on the architectural vision and framework set by the first version of the white paper, the Version 2.0 of the white paper was released in January 2018 and presented the latest findings and analyses of 5G PPP Phase I projects along with the concept evaluations.

The work has continued with the 5G PPP Phase II and Phase III projects with special focus on understanding the requirements from vertical industries involved in the projects and then driving the required enhancements of the 5G Architecture able to meet their requirements. The results of the Working Group are now captured in this Version 3.0, which presents the consolidated European view on the architecture design.

Table of Contents

1	Introduction.....	25
2	Overall architecture.....	26
2.1	Stakeholder roles in the 5G ecosystem.....	26
2.2	5G Enhanced Overall System Architecture	27
2.3	E2E Service Operations – Lifecycle Management	29
2.4	Domain Management & Orchestration.....	31
2.4.1	Multi-domain management.....	31
2.4.2	Application-aware orchestration.....	32
2.4.3	Service-specific extensions of ETSI NFV MANO.....	33
2.5	Programmable Networks	33
2.5.1	Data plane programmability	34
2.5.2	Transport network programmability	35
2.5.3	Network function programmability in RAN.....	36
2.6	Vertical-specific architecture extensions	37
2.6.1	Extensions for energy utilities	37
2.6.2	Extensions for vehicular communications.....	38
2.6.3	Extensions for enhanced content delivery	39
2.6.4	Extensions for media production and delivery	40
2.7	References.....	41
3	Radio & Edge Architecture	43
3.1	Overall RAN Architecture	43
3.1.1	Centralized-Distributed-Radio Units split and Control-User Planes separation.....	44
3.2	Protocol Extensions for Vertical Support	45
3.2.1	RAN Part of Network Slicing.....	45
3.2.1.1	Fundamental Slicing Support in RAN.....	45
3.2.1.2	Wi-Fi virtualization and slicing.....	46
3.2.2	RAN Analytics	48
3.2.3	Virtualized Small Cell (Cloud-enabled small cell) at the Light Data Centre	49
3.2.4	Multi-access Edge Computing & Energy Efficient Radio Access Networks	51
3.2.5	Multi-link Cooperation	53
3.2.6	Local end-to-end Path.....	55
3.2.7	Multicast and Broadcast in RAN.....	56
3.2.8	Non-terrestrial networks (NTN)	59
3.3	RAN Architecture Implementation.....	63
3.3.1	RAN Deployment Options	63
3.3.2	Visible Light Communication-based gNB	65

3.3.3	Lower Layer Split and Implementation	68
3.3.3.1	Baseband processing in active antenna distributed unit	68
3.3.3.2	Optical Beamforming	69
3.4	Edge Architecture	73
3.4.1	XMEC: extension of edge computing functionalities.....	73
3.5	References.....	74
4	Core & Transport Architecture	76
4.1	Introduction.....	76
4.2	Core Network Architecture.....	76
4.2.1	Multicast and Broadcast	77
4.2.2	Analytics Framework.....	79
4.2.2.1	Data Analytics Characterization.....	80
4.2.2.2	Integrated Analytics Architecture.....	81
4.3	Transport network Infrastructure	83
4.3.1	Wireline technologies	83
4.3.1.1	Programmable Elastic frame-based Optical Transport.....	83
4.3.1.2	Ethernet Transport	85
4.3.1.3	Programmable Metro Network - Disaggregated Edge Node.....	87
4.3.1.4	Space Division Multiplexing	88
4.3.2	Wireless	90
4.3.2.1	Millimeter wave (mmWave)	91
4.3.2.2	Multi-tenant small cells with IAB	92
4.3.2.3	Satellite Backhaul.....	95
4.3.3	Fiber-Wireless (FiWi) PtMP.....	97
4.4	Examples of data plane programmability	100
4.4.1	Programmable Data Plane	100
4.4.2	Stateful Packet Processing in Hardware	101
4.4.3	Segment routing.....	102
4.5	References.....	103
5	Management & Orchestration Architecture.....	106
5.1	Introduction.....	106
5.2	Starting point: Existing High-level MANO architectures	106
5.2.1	ETSI summary	106
5.2.1.1	ETSI NFV & MANO	106
5.2.1.2	ETSI ZSM	106
5.2.1.3	ETSI ENI.....	107
5.2.1.4	ETSI MEC	107
5.2.2	3GPP	107

5.2.3	Status and consensus architecture.....	108
5.2.3.1	Example aspect: Heterogeneity	110
5.3	Architecture options beyond the consensus architecture	112
5.3.1	Structure of orchestration	112
5.3.1.1	Service-type: Integrated or segregated orchestration	112
5.3.1.2	Flat vs. hierarchical orchestration.....	113
5.3.1.3	Orchestration vs. slicing	114
5.3.1.4	Abstractions and their violations	115
5.3.1.5	Conflict resolution	116
5.3.1.6	Time scales	116
5.3.1.7	Technologies.....	117
5.3.2	Implementation patterns for MANO frameworks	118
5.3.3	Algorithmic building blocks	119
5.3.3.1	Example building blocks	119
5.3.3.2	Realizations, MAPE-K example.....	120
5.3.4	Description mappings	121
5.3.4.1	Horizontal mapping	121
5.3.4.2	Vertical mapping	122
5.3.5	Monitoring aspects in Orchestration.....	122
5.4	DevOps meets Orchestration	123
5.4.1	Development time: SDKs	123
5.4.2	Need for validation tools	124
5.4.3	Need for evaluation tools.....	124
5.5	References.....	125
6	Deployment, Evaluations & Analyses	127
6.1	Introduction.....	127
6.1.1	Performance KPIs.....	127
6.1.2	Types of evaluation and usability of KPIs.....	128
6.1.3	Approach/options to generation of KPIs	129
6.1.4	Architecture instantiations	130
6.2	Management and orchestration.....	134
6.3	Verification and Validation System.....	136
6.4	Emulation Framework	137
6.5	References.....	138
7	Impact on standardization	139
7.1	Brief profile of targeted SDOs.....	139
7.1.1	3GPP	139
7.1.2	ETSI.....	140

7.1.3	ITU.....	140
7.1.4	IETF.....	141
7.1.5	Other SDOs.....	141
7.2	Standards impact by 5G PPP Phase 2 projects	142
7.2.1	Contributions related to overall architectures	143
7.2.2	Contributions related to radio and edge architectures	144
7.2.3	Contributions related to core and transport architectures	145
7.2.4	Contributions related to management and orchestration architectures	146
7.3	Fostering Standardization via 5G PPP.....	147
7.4	References.....	147
8	Conclusions and Outlook	149
9	Glossary	152
9.1	General Terms	152
9.2	Network function virtualization related	152
9.3	Network slice related	154
9.4	Vertical service related	154
9.5	Multi-access edge computing related	155
9.6	Business logic/stakeholder related.....	156
9.7	DevOps-related terms and roles.....	156
9.8	Specific terms	159
9.9	References.....	161
10	List of Contributors.....	164
11	List of 5G PPP Phase 2 Contributions to Standards	169

List of Acronyms and Abbreviations

3GPP	3rd Generation Partnership Project
3GPP-CT1	3GPP TSG-CT WG1 MM/CC/SM (Iu)
3GPP-CT3	3GPP TSG-CT WG3 Interworking with external networks
3GPP-CT4	3GPP TSG-CT WG4 MAP/GTP / BCH/SS
3GPP-RAN1	3GPP TSG-RAN WG1 Radio Layer 1
3GPP-RAN2	3GPP TSG-RAN WG2 Radio Layer 2 and Radio Layer 3 RR
3GPP-RAN3	3GPP TSG-RAN WG3 Iu, Iub, Iur, S1, X2 and UTRAN/E-UTRAN
3GPP-SA1	3GPP TSG-SA WG1 Services
3GPP-SA2	3GPP TSG-SA WG2 Architecture
3GPP-SA3	3GPP TSG-SA WG3 Security
3GPP-SA4	3GPP TSG-SA WG4 Codec
3GPP-SA5	3GPP TSG-SA WG5 Telecom Management
3GPP-SA6	3GPP TSG-SA WG6 Mission-critical applications
5GAA	5G Automotive Association
5G-ACIA	5G Alliance for Connected Industries and Automation
5G CAR	Fifth Generation Communication Automotive Research and innovation
5G City	<i>no long form</i>
5G ESSENCE	Embedded Network Services for 5G Experiences
5G IA	5G Infrastructure Association
5G TANGO	5G Development and Validation Platform for global Industry – specific Network Services and Apps
5G TRANSFORMER	5G Mobile Transport Platform for Verticals
5G XCAST	Broadcast and Multicast Communication Enablers for the Fifth Generation of Wireless Systems
5GC	5G Core
5G-MoNArch	5G Mobile Network Architecture for diverse services, use cases, and applications in 5G and beyond

5G-PICTURE	5G Programmable Infrastructure Converging disaggregated neTwork and compUte Resources
5G PPP	5G Infrastructure Public Private Partnership
5GS	5G System
AADU	Active Antenna Distributed Unit
AC	All-Connected
ACTN	Abstraction and Control of Traffic engineered Networks
AF	Application Function
AL-DAF	Application Layer Data Analytics Function
AL-FEC	Application-Level Forward Erasure Correction
AL-FEC	Application Layer FEC
AMF	core access and Mobility management Function
AP	Access Point
API	Application Programming Interface
APP	Application
ARoF	Analogue Radio over Fibre
ARP	Allocations and Retention Priority
ASN	Abstract Syntax Notation
BB	Baseband
BBU	Baseband Unit
BCH	Broadcast Channel
BH	Backhaul
blueSpace	Building on the Use of Spatial Multiplexing 5G Networks Infrastructures and Showcasing Advanced technologies and Networking Capabilities
BoF	Birds of a Feather
BS	Base Station
BSS	Business Support System
CA	Carrier Aggregation
CAPEX	Capital Expenditure

CAPIF	Common API Framework
CAS	Cell Acquisition Subframe
CBR	Constant bitrate
CBTR	Coding-Building-Testing-Releasing
CC	Call Control
CCAMP	Common Control and Measurement Plane
CD	Continuous Delivery
CESC	Cloud Enabled Small Cell
CESCM	Cloud Enabled Small Cell Manager
C-ITS	Cooperative Intelligent Transportation System
CN	Core Network
CO	Central Office
CoMP	Coordinated Multi-Point
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CPU	Central Processing Unit
CQI	Channel Quality Indicator
CRAN	Centralised RAN
cRRM	Centralised RRM
CS	Cell site
CSAR	Cloud Service Archive
CSDPS	Coordinated Scheduling and Dynamic Point Selection
cSD-RAN	centralized Software-Defined Radio Access Network (cSD-RAN)
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
cSON	centralised SON
CT	Core Network & Terminals
CU	Centralised Unit
DAF	Data Analytics Function

DASH	Dynamic Adaptive Streaming over HTTP
DC	Dual Connectivity
DCSP	Data Centre Service Provider
DFT	Discrete Fourier Transform
DL	Downlink
DN	Data Network
DN-DAF	Data Network Data Analytics Function
DPDK	Data Plane Development Kit
DPI	Deep Packet Inspection
DR	Demand Response
DRAN	Distributed RAN
DRB	Data Radio Bearer
DRoF	Digital Radio over Fibre
dRRM	distributed RRM
DSC	Digital Service Customer
dSON	distributed SON
DSP	Digital Service Provider
DU	Distributed Unit
DVB	Digital Video Broadcasting
DVB TM-IPI	DVB Technical Module - Internet Protocol Infrastructure
E2E	End-to-End
eCPRI	Enhanced CPRI
eICIC	Enhanced ICIC
eMBB	enhanced Mobile Broadband
eNA	Enablers for Network Automation for 5G
eNB	evolved Node B
ENI	Experiential Networked Intelligence
EoMPLS	Ethernet over Multi-Protocol Label Switching

EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
eV2X	Enhancement of V2X
EVM	Error Vector Magnitude
FaaS	Function-as-a-Service
FAPI	Functional Application Platform Interface
FDD	Frequency Division Duplex
FDT	File Description Table
FEC	Forward Error Correction
FFS	For Further Study
FFT	Fast Fourier Transform
FH	Fronthaul
FiWi	Fiber-Wireless
FlexO-LR	Flexible OTN long-reach
FPGA	Field Programmable Gate Array
FRASE	FEC and ROHC Activation for GCSE over MBMS
FS-eNA	Study of enablers for Network Automation for 5G
FSM	Function-Specific Manager
GCSE	Group Communication System Enablers
GEO	Geostationary
GFDM	Generalised Frequency Division Multiplexing
GLPK	GNU Linear Programming Kit
gNB	Next Generation Node B
gNB-CU-MC	gNB-centralized unit-multicast
GPP	General Purpose Platform
GPRS	General Packet Radio Service
GPU	Graphics Processing Unit
G-RNTI	Group RNTI

GST	Guaranteed Service Transport
GTP	GPRS Tunnelling Protocol
GTP-U	GTP-User Plane
GUI	Graphical User Interface
HAPS	High Altitude Platform Station
HARQ	Hybrid Automatic Repeat Request
HEO	Highly Elliptical Orbit
HPHT	High-Power High-Tower
HTTP	Hypertext Transfer Protocol
HW	Hardware
IaaS	Infrastructure as a Service
IAB	Integrated Access and Backhaul
ICIC	Inter Cell Interference Coordination
IE	Information Element
IETF	Internet Engineering Task Force
IFA	Interfaces and Architecture
IFFT	Inverse FFT
IGMP	Internet Group Management Protocol
IHPCGW	Intelligent Home IP Gateway
IIoT	Industrial Internet of Things
IMS	IP-Multimedia Subsystem
IOC	Information Object Class
IoRL	Internet of Radio Light
IoT	Internet of Things
IRP	Integration Reference Point
ISC	Intra-Slice Controller
ISD	Inter-Site Distance
ISG	Industry Specification Group

ISU	Interface Sub-Unit
ITU	International Telecommunication Union
ITU-D	ITU Development sector
ITU-R	International Telecommunication Union Radiocommunication Sector
ITU-T	International Telecommunication Union Telecommunication Standardization Sector
IWF	InterWorking Function
JR	Joint Reception
KPI	Key Performance Indicator
KPI	Key Performance Indicator
LAN	Local Area Network
LCM	Lifecycle Management
LEO	Low Earth Orbit
LLC	Logical Link Control
LoA	Level of Automation
LPWA	Low Power IoT Networking
LS	Liaison Specification
LTE	Long Term Evolution
LwM2M	Lightweight Machine to Machine
M2M	Machine-to-Machine
MAC	Medium Access Control
MANO	Management and Orchestration
MAP	Mobile Application Part
MAPE-K	Monitor-Analyze-Plan-Execute over shared Knowledge
MATILDA	A Holistic, Innovative Framework for the Design, Development and Orchestration of 5G-ready Applications and Network Services over Sliced Programmable Infrastructure
MBB	Mobile Broadband
MBMS	Multimedia Broadcast/Multicast Service

Mbps	Megabits per second
MC	Multi-Connectivity
MC-PTM	Multi-cell point-to-multipoint
MCDData	Mission Critical Data
MCF	Multi-Core Fibre
MCPTT	Mission Critical Push To Talk
MCS	Modulation and Coding Scheme
MCX	Mission Critical Services
MDAF	Management and Orchestration Data Analytics Function
MEC	Multi-Access Edge Computing
MEF	Metro Ethernet Forum
MEO	Medium Earth Orbit
METRO-HAUL	METRO High bandwidth, 5G Application-aware optical network, with edge storage, compUte and low Latency
MFN	Multiple Frequency Network
MIMO	Multiple-Input Multiple-Output
MM	Mobility Management
MME	Mobility Management Entity
mMTC	massive Machine-Type Communication
mmWave	Millimetre Wave
MNO	Mobile Network Operator
MNO	Mobile Network Operator
MOCN	Multi-Operator Core Network
M-OTN	Mobile-optimized OTN
MPTCP	Multipath TCP
MSA	Multi-Source Agreement
MTO	Multi-Tier Orchestrator
MulteFire	<i>no long form</i>

NAS	Non Access Stratum
NAT	Network Address Translation
NBI	Northbound Interface
NEF	Network Exposure Function
NETCONF	Network Configuration protocol
NF	Network Function
nFAPI	network FAPI
NFV	Network Functions Virtualization
NFVI	Network Function Virtualisation Infrastructure
NFVI	NFV Infrastructure
NFV-NSI	NFV Network Service Information
NFVO	NFV Orchestrator
NGFI	Next Generation Fronthaul Interface
NGMN	Next Generation Mobile Networks
NGPaaS	Next Generation Platform as a Service
NG-RAN	Next Generation RAN
NIC	Network Interface Card
NLoS	Non-Line-of-Sight
NOP	Network Operator
NR	New Radio
NRE	Near Real-time Engine
NRF	Network Repository Function
NRG-5	Enabling Smart Energy as a Service via 5G Mobile Network advances
NRM	Network Resource Model
NS	Network Slice
NSA	Non-Standalone
NSaaS	Network Slice as a Service
NSD	Network Service Descriptor

NSI	Network Slice Instance
NSO	Network Service Orchestration
NSSAI	Network Slice Selection Assistance Information
NW	Network
NTN	Non-terrestrial networks
NWDA	Network Data Analytics
NWDAF	NWDA Function
NWMO	Network Management and Orchestration
OAM	Operations Administration and Maintenance
OBFN	Optical Beam Forming Network
OBSAI	Open Base Station Architecture Initiative
ODL	Open Daylight
ODN	Optical Distribution Network
OFDM	Orthogonal Frequency Division Multiplexing
ONF	Open Network Foundation
OPEX	Operational Expenditure
OPNFV	Open Platform for NFV
OS	Operating System
OSI	Open Systems Interconnection
OSM	Open Source MANO
OSNR	Optical Signal-to-Noise Ratio
OSS	Operations Support System
OTDR	Optical Time Domain Reflectometer
OTN	Optical Transport Network
OTU4	Optical Channel Transport Unit 4
OvS	Open virtual Switch
OvSDB	Open vSwitch Database
pCR	pseudo Change Request

PCS	Physical Coding Sublayer
PDB	Packet Delay Budget
PDU	Protocol Data Unit
PDV	Packet Delay Variation
PER	Packet Error Rate
PHY	Physical (layer)
PLMN	Public Land Mobile Network
PM	Performance Management
PNF	Physical Network Function
PNFM	Physical Network Function Manager
PoC	Proof of Concept
PON	Passive Optical Network
PoP	Point-of-Presence
PPP	Public Private Partnership
PRACH	Physical Random Access Channel
PSCE	Public Safety Communications Europe
PTM	Point-to-Multipoint
PTP	Point-to-Point
PTP	Precision Time Protocol
PWS	Public Warning System
QoE	Quality of Experience
QoS	Quality of Service
QoS	Quality of Service
QUIC	<i>no long form</i>
RACH	Random Access Channel
RAN	Radio Access Network
RAN-DAF	Radio Access Network Data Analytics Function
RAT	Radio Access Technology

RCA	RAN Controller Agent
REST	REpresentational State Transfer
RF	Radio Frequency
RFC	Request For Comment
RLC	Radio Link Control
RLV	Radio Link Control
RM	Resource Management
RMA	RAN Multicast Area
RN	Relay Node
RNTI	Radio Network Identifier
ROADM	Reconfigurable Optical Add/Drop Multiplexers
ROHC	Robust Header Compression
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RS-U	Radio Sub-Units
RSU	Road Side Unit
RU	Radio Unit
RU	Radio Unit
SA	Service & Systems Aspects
SaaS	Software as a Service
SAND	Server And Network-assisted DASH
SAT5G	Satellite and Terrestrial Network for 5G
SBA	Service-Based Architecture
SBI	Service Based Interface
SC	Service Customer
SDAP	Service Data Adaptation Protocol
SDK	Service Development Kit

SDK	Software Development Kit
SDM	Space Division Multiplexing
SDN	Software-Defined Networking
SDO	Standards Developing Organisation
SES	Satellite Earth Stations and Systems
SFC	Service Function Chaining
SFN	Single Frequency Networks
SG	Study Group
SGW	Serving Gateway
SI	Study Item (3GPP)
SIC	Successive Interference Cancellation
SID	Shared Information Data
SL	Sidelink
SLA	Service-Level Agreement
SLICENET	End-to-End Cognitive Network Slicing and Slice Management Framework in Virtualised Multi-Domain, Multi-Tenant 5G Networks
SLO	Service Level Objective
SM	Session Management
SME	Small and Medium-sized Enterprise
SMF	Session Management Function
S-NSSAI	Single NSSAI
SoBI	Southbound Interface
SON	Self-Organising Network
SP	Service Provider
SR	Segment Routing
SRB	Signalling Radio Bearer
SRI	Satellite Radio Interface
SSID	Service Set Identifier

SSM	Service-Specific Manager
SW	Software
TAPI	Transport API
TCP	Transmission Control Protocol
TD	Time Division
TDD	Time Division Duplex
TDM	Time Division Multiplexing
TE	Traffic Engineering
TN	Transport Network
TOSCA	Topology and Orchestration Specification for Cloud Applications
TR	Technical Report (ETSI)
TS	Technical Specification (ETSI)
TSDR	Time Series Data Repository
TSG	Technical Specification Group
TSG-CT	TSG Core Network and Terminals
TSG-RAN	TSG Radio Access Networks
TSG-SA	TSG Service and System Aspects
TSN	Time Sensitive Networking
TSON	Time Shared Optical Network
TTI	Transmission Time Interval
TTM	Time-To-Market
UAS	Unmanned Aircraft System
UE	User Equipment
UPF	User Plane Function
URLLC	Ultra-Reliable Low-Latency Communication
USIM	Universal Subscriber Identity Module
UTRAN	Universal Terrestrial Radio Access Network
V2N	Vehicle-to-Network

V2P	Vehicle-to-Pedestrian
V2X	Vehicle-to-Everything
vAAA	virtual Authentication, Authorization, Accounting
vaps	virtual Access Points
vBBU	Virtual Baseband Unit
vBCP	Virtual Blockchain Processing
vDES	virtual Distributed Energy Storage
vDFC	virtual Drone Flight Control
vESR	virtual Electricity Substation & Rerouting
V-GW	Virtual Gateway
VIF	Virtual Interface
VIM	Virtualised Infrastructure Manager
VISP	Virtualisation Infrastructure Service Provider
VLAN	Virtual LAN
VLC	Visible Light Communication
VLIW	Very Long Instruction Word
VM	Virtual Machine
vMCM	virtual Machine-Cloud-Machine
vMME	Virtual MME
vMPA	virtual Media Processing & Analysis
VN	Virtual Network
VNE	Virtual Network Element
VNF	Virtualized Network Function
VNFD	Virtualized Network Function Descriptor
VNFM	VNF Manager
VNN	Vehicular Nomadic Node
vPMU	virtual Phasor Measurement Unit
VPN	Virtual Private Network

VPON	Virtual PON
vRES	virtual Renewable Energy Sources
VSB	Vertical Service Blueprint
VSD	Vertical Service Descriptor
VSD	Virtual Service Descriptor
vSON	Virtual SON
vTSD	virtual Terminals Self-Discovery
WAN	Wide-Area Network
WDM	Wavelength Division Multiplexing
WDM	Wavelength Division Multiplex
WEP	Wired Equivalent Privacy
WG	Work Group
WI	Work Item (3GPP)
WID	Work Item Description
WIM	WAN Infrastructure Manager
WLAN	Wireless Local Area Network
WPA	Wi-Fi Protected Access
WSS	Wavelength Selective Switch
XCF	Xcast Control Function
xMB API	extended MBMS Interface
XMEC	Extended MEC
XML	eXtensible Markup Language
XSC	Cross-Slice Controller
XUF	Xcast User Plane Function
YANG	Yet Another Next Generation
ZSM	Zero touch network and Service Management

1 Introduction

5G is the first generation of mobile networks that is by design addressing the wide range of needs of the vertical industries. 5G offers unlimited mobile broadband experience, provides massive connectivity for everything from human-held smart devices to sensors and machines, and most importantly, it has the ability to support critical machine communications with instant action and ultra-high reliability.

First 5G specifications are available with 3GPP Rel.15 where the focus was primarily to serve mobile operator needs in terms of extreme mobile broadband services. A second release (3GPP Rel.16) will be soon available end of 2019 that will include several features to support vertical in terms of enablers for Industrial IoT and URLLC. However, this is only the initial step, further enhancements and optimizations are still needed to design a 5G System that meets the challenging requirements from the vertical industries. With this White Paper the 5GPPP Architecture Working Group is going to summarize the finding from 5GPPP Phase 2 and Phase 3 projects with the aim to impact the coming standards releases.

The White Paper is organized as follows. The overall architecture in Chapter 2 provides several enhancements to the 3GPP Rel.15 system architecture to address specific requirements from vertical industries. Chapter 3 details the RAN architecture and the relevance of the edge to enable local computing and local path to support diverse variety of requirements in terms of latency, throughput and reliability. Chapter 4 complements previous chapter describing the CN architecture and the infrastructure connectivity provided by the Transport Network architecture. In order to achieve the required flexibility, the management and orchestration architecture plays a key role as presented in Chapter 5. To enable different verticals and large variety of requirements need to be supported, in Chapter 6 we present examples of architecture deployments investigated and analysed in the 5GPPP projects. Finally, Chapter 7 summarizes the standardization impact that the work done have already achieved in different standardization bodies with special focus on 3GPP and ETSI.

2 Overall architecture

5G networks have been targeted to meet the requirements of a highly mobile and fully connected society. The coexistence of human-centric and machine type applications will define very diverse functional and performance requirements that 5G networks will have to support. Within the 5G System (5GS), end-to-end (E2E) network slicing, service-based architecture, Software-Defined Networking (SDN), and Network Functions Virtualisation (NFV) are seen as the fundamental pillars to support the heterogeneous key performance indicators (KPIs) of the new use cases in a cost-efficient way. The 5GS gives mobile network operators the unique opportunities to offer new services to consumers, enterprises, verticals, and third-party tenants by addressing their respective requirements. To this end, 5G Infrastructure Public Private Partnership (5G PPP) Phase I/II collaborative research projects as well as standardisation bodies have specified and developed the main elements of the 5G architecture.

2.1 Stakeholder roles in the 5G ecosystem

The 5G ecosystem should enable manufacturers, solution integrators, network and service providers, and Small and Medium-sized Enterprises (SMEs) to efficiently compete and cooperate, e.g., by means of virtualisation, standardised interfaces and protocols, or open APIs. SMEs will be able to provide technological solutions which will be compatible with the overall system, e.g., new hardware components in the infrastructure or software components in the Management and Organization layers. Manufacturers and solution integrators can offer rapid deployment enabled by virtualisation and standardised interfaces to increase the level of innovation. Mobile Network Operators (MNOs) and infrastructure providers will create tailored slices with specific functionalities and Over-The-Top applications and services to address requirements of vertical industries.

[2-1] and [2-5] mainly focus on single-domain service provisioning, but do not elaborate on, e.g., cross-operator scenarios. Additionally, the 3GPP roles are defined from the point of view of an operator. 5G PPP Phase II and III collaborative research projects have extended these roles to allow various possible customer-provider relationships between verticals, operators, and other stakeholder, as shown in Figure 2-1. This figure is a consolidated and agreed 5G stakeholder role model among several projects (e.g. [2-6], [2-7], [2-8]):

- **Service Customer (SC):** uses services that are offered by a Service Provider (SP). In the context of 5G, vertical industries are considered as one of the major SCs.
- **Service Provider (SP):** comprises three sub-roles, depending on the service offered to the SC: Communication Service Provider offering traditional telecom services, Digital Service Provider offering digital services such as enhanced mobile broadband and IoT to various vertical industries, or Network Slice as a Service (NSaaS) Provider offering a network slice along with the services that it may support and configure. SPs design, build and operate services using aggregated network services.
- **Network Operator (NOP):** in charge of orchestrating resources, potentially from multiple virtualised infrastructure providers (VISP). The NOP uses aggregated virtualised infrastructure services to design, build, and operate network services that are offered to SPs.
- **Virtualisation Infrastructure Service Provider (VISP):** Provides virtualised infrastructure services and designs, builds, and operates virtualisation infrastructure(s) [2-1]. The infrastructure comprises networking (e.g., for mobile transport) and computing resources (e.g., from computing platforms).
- **Data Centre Service Provider (DCSP):** Provides data centre services and designs, builds and operates its data centres. A DCSP differs from a VISP by offering “raw” resources (i.e., host servers) in rather centralised locations and simple services for consumption of these raw resources. A VISP rather offers access to a variety of resources by aggregating multiple technology domains and making them accessible through a single API.

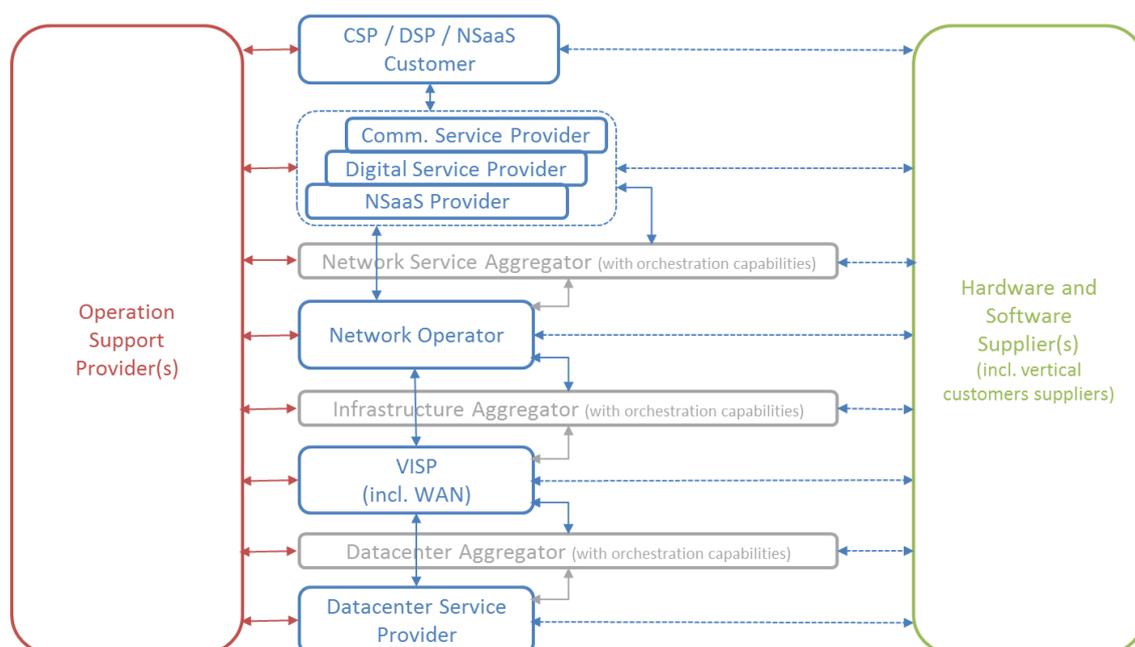


Figure 2-1: Stakeholder roles in the 5G ecosystem

2.2 5G Enhanced Overall System Architecture

The prospects of network slicing, i.e., executing multiple logical mobile network instances on a shared infrastructure, require a continuous reconciliation of customer-centric service level agreements (SLAs) with infrastructure-level network performance capabilities. Service customers, e.g., from the vertical industries, request the creation of (tele)communication services by providing “customer-facing” on-demand service requirement descriptions to Service Providers. In the past, operators executed such mapping in a manual manner on a limited number of service/slice types (mainly mobile broadband, voice, and SMS). With an increased number of such customer requests, an E2E framework for **Service Creations** and Service Operations will therefore have to exhibit a significantly increased level of automation for the lifecycle management of network slice instances.

On the **Service Level**, lifecycle management automation must be realized by closed-loop Service Assurance, Service Fulfilment, and Service Orchestration functions (cf. Figure 2-2) covering all lifecycle phases: preparation phase, instantiation, configuration and activation phase, run-time phase, and decommissioning phase. Two fundamental technological enablers include softwarisation, e.g., virtualisation of network functions, as well as software-defined, programmable network functions and infrastructure resources. **E2E Service Operations** functions interact with functions for **Management of Domain Resources and Functions**. Example domains include RAN, Core & Transport Network, as well as NFV and MEC. Besides orchestration, closed-loop procedures for resource fulfilment, resource assurance, and network intelligence comprise building blocks within each management domain. On a more fine-grained temporal and spatial level, domain-specific controllers, incl. SDN controllers, can be programmed to efficiently execute policies and rules on the **Resources and Functional Level**.

Finally, a common platform, where data can be accessed by system entities from all levels, uses scalable data exposure governance and access control mechanisms to provide services for data acquisition, processing, abstraction, and distribution. This includes data related to subscribers, to the network and underlying resources, to network slice and service instances, and, if required by the vertical customer, to applications.

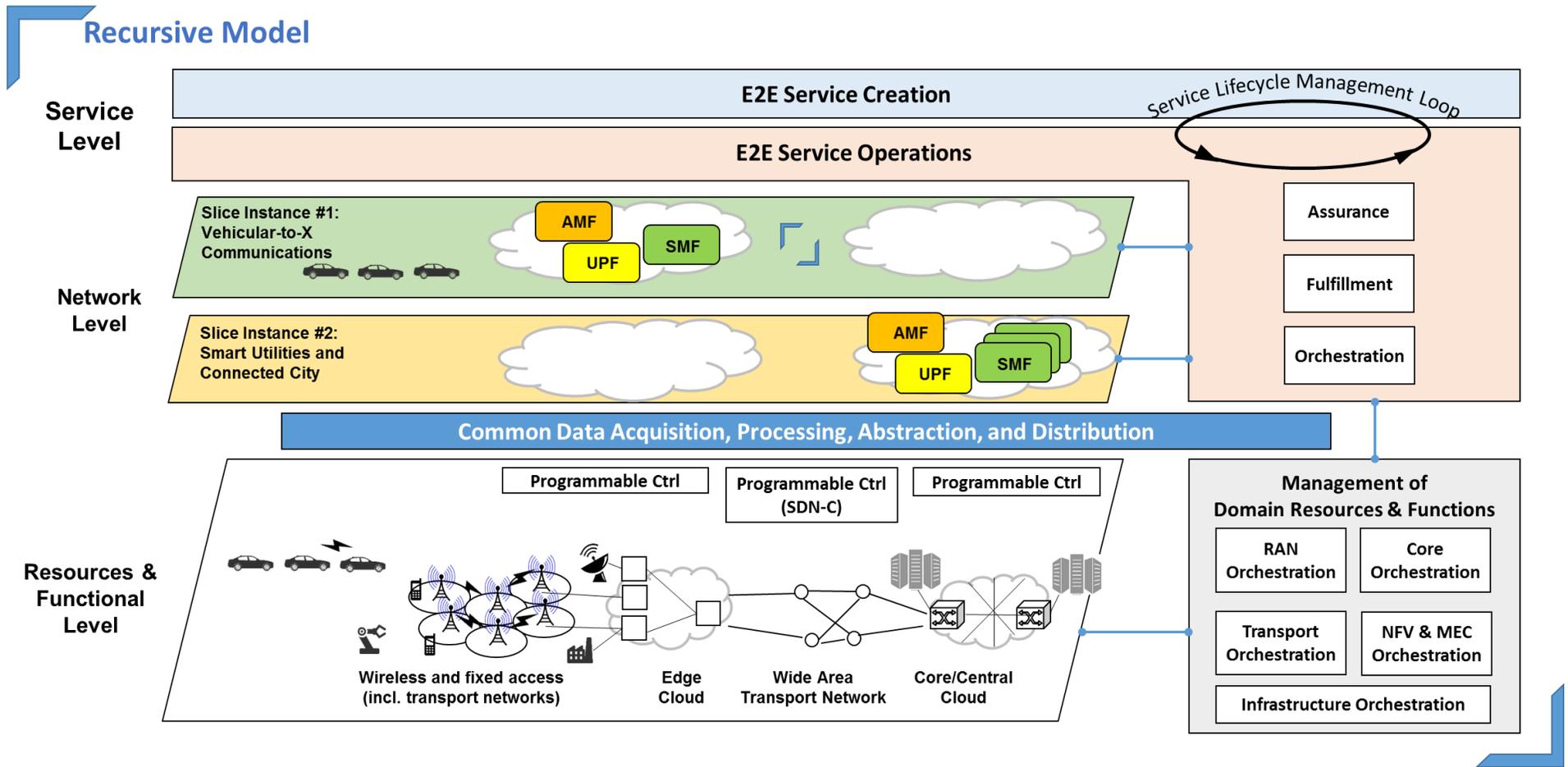


Figure 2-2: Overall Architecture

The proposed architecture realizes a recursive structure. A recursive structure in the 5G context can be defined as a design, rule, or procedure that can be applied repeatedly. In a network service context, this recursive structure applies to a specific part of either a network service or the deployment platform. It is defined as the ability to build a service out of existing services, including another instance of the very same service. As with a recursive service definition, a recursive structure in the 5G architecture can be instantiated and linked repeatedly. It improves scalability, as the same service category can be deployed many times, at different places at the same time. Delegating parts of the service to multiple instances of the same software block is a natural way to handle more complex and larger workloads or service graphs.

In the context of virtualised infrastructure, such recursive structure allows a slice instance operating on top of the infrastructure resources provided by another slice instance. For example, each tenant can own and deploy its own Management and Orchestration (MANO) system. To support the recursion, a set of homogeneous APIs are needed for providing a layer of abstraction for the management of each slice and controlling the underlying virtual resources which is transparent to the level of the hierarchy where the tenant is operating. Different tenants request the provisioning of slices through these APIs. By means of a template, blueprint, or SLA, each tenant specifies not only the slice characteristics (topology, QoS, etc.) but also some extended attributes such as the level of resiliency, management and control desired.

2.3 E2E Service Operations – Lifecycle Management

In 5G, following the cloud software platform-infrastructure model [2-20], many things will be offered as a service, including infrastructure, a platform, or software. The concept of network slicing is expected to satisfy the need for customised, service-specific combinations of service components and network functions in all of the network segments. Service lifecycle management (LCM) tools are enabled by Service Development Kits (SDKs). Using SDKs, services can be reconfigured, or new service versions can be created. As shown in Figure 2-3, the service lifecycle starts with service development. After ensuring that Quality of Service (QoS) expectations of the end-users are met, it can be deployed to the production environment. The deployed services continue to be monitored during the operations stage which completes the lifecycle. To optimize the service development with QoS expectations, a variety of SDK approaches are proposed by several H2020 projects, among them [2-10], [2-11], [2-14], and [2-15].

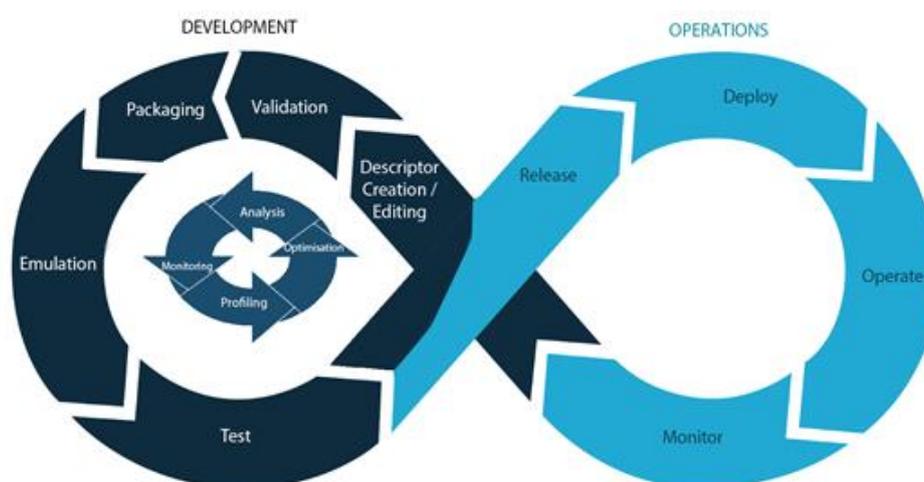


Figure 2-3: The lifecycle of a service

Table 2-1 denotes tools and functionalities of the SDK for end-to-end service LCM.

Table 2-1: SDK tool and functionality for end-to-end service LCM

Descriptor Creation Tools	A Network Service Descriptor (NSD) is a deployment template that contains the information used by the Network Functions Virtualisation (NFV) Orchestrator (NFVO) for lifecycle management of a Network Service (NS). Similarly, a Virtualised Network Function Descriptor (VNFD) defines the resources required for realising a Virtualised Network Function (VNF). The creation of NSD and VNFD based on ETSI NFV Industry Specification Group (ISG) (ETSI, 2019) follows a two-step approach: (i) The specification of language-independent information models that identify the content of the descriptors using information elements. (ii) Translation of these information models into language-specific data models and package formats, e.g., [2-18], [2-19].
Descriptor Validators	The application languages may fail to describe a complete NS as they usually lack a detailed description of network functionality and requirements. Therefore, in the development stage, the VNFDs and NSDs should be validated against a given schema before being processed by the dispatch engine
Packaging Tools	The packaging tool defines a standard way for VNF providers to deliver VNFs to service providers. The VNF package presents the complete picture of the VNF, by combining parameters from the descriptors, the image to be used for it and the configuration that it should be deployed with
NFVI Emulators	The emulator facilitates local-machine prototyping and testing of NSs in realistic end-to-end, multi-PoP (point of presence) topologies and multi-VIM environments.
Profiling Tools	The profiling tool provides load testing under various resource constraints on the NSs by monitoring a variety of metrics. E.g., service developers can detect bottlenecks, realise resource and elasticity profiling, or evaluate if an NS ensures the expected QoS.
Optimization Tools	If an NS do not meet the QoS expectations, the optimization tools may optimize the NS, e.g., based on machine learning algorithms.
Deployment Tools	After confirming that the QoS expectations are met, the verified NSs are deployed in the production environment using the deployment tools.
Monitoring Tools	Containerisation and virtualisation techniques require different kind of NS customisations and configurations. To avoid bugs and decrease the time required between the development process and the operations of an application, monitoring tools allow rapid testing and verification of any modified parameter.

Descriptor Creation Tools can also be used to create and customise network service descriptors. This allows for the definition of vertical service blueprints (VSB) with specific SLA requirements on the services. VSBs may contain parameters of different types, e.g. IP addresses of external components, numbers of supported users, and even additional VNF packages [2-9]. An industry vertical (or other customers) can provide actual values for these parameters, thereby creating a customised vertical service descriptor (VSD). SLA requirements such as latency on paths,

availability of the service, or energy-efficient service deployment, facilitate the selection of specific NSDs, their deployment flavours, and instantiation levels.

2.4 Domain Management & Orchestration

2.4.1 Multi-domain management

E2E services have to be realised by resources and functions from multiple network, technology, and administrative domains, cf. Figure 2-2. This requires interaction between the E2E Service Operations and each of the involved management domains.

For classification of such multi-domain scenarios, [2-5] defines two general scenarios where services are provided across multiple service providers, i.e., administrative domains:

1. Classical roaming: Individual users move from one provider (i.e. home network provider) to a network of another provider (i.e. visited network provider). The services that a user requires while roaming needs to be specified in the SLA between the two providers.
2. Business verticals scenario: a business vertical user's request may only be met by the capabilities of multiple service providers. The user's main service provider therefore has to extend its own capabilities with capabilities from another service provider by means of according SLAs.

[2-12] further defines the high-level mechanisms for multi-domain configuration, i.e., if two or more different administrative domains are required to cooperate to provide the necessary resources and functions to support any given service. The network slice required to support the service is established through a cooperation of the domain-specific orchestrators, based on policies and agreements that are applicable across the participating administrative domains.

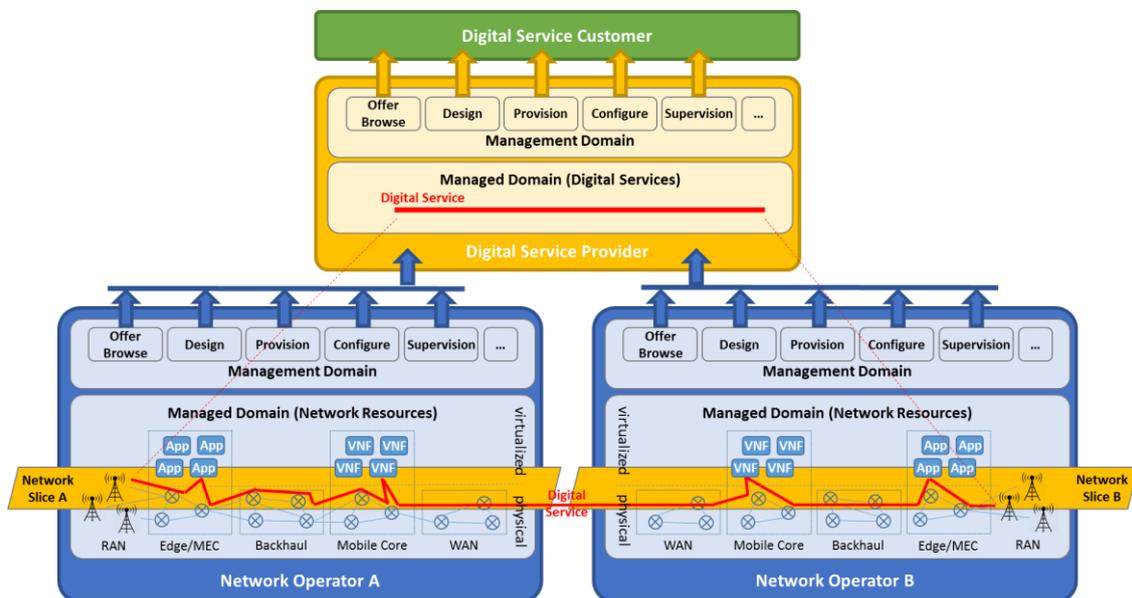


Figure 2-4: Multi-domain management for e2e service composition

Figure 2-4 describes a Digital Service Customer (DSC) requesting a digital service from the Digital Service Provider (DSP). The composition of the service requires network resources and functions from NOPs A and B. Both NOPs manage their network infrastructures and expose them to the DSP. The DSP has to browse, select, manage, and orchestrate the network services from different NOPs in order to create a composed digital service based on the needs of the DSC. Moreover, the DSP will implicitly consume infrastructure services that both NOPs aggregate to compose their respective network service. Finally, the DSC will consume the composed digital

service without the need to know that it is composed by services of several NOPs. I.e., the DSC unaware about network slices instances A and B that are used within the NOP domain to deliver the digital service of the DSP.

2.4.2 Application-aware orchestration

In accordance with the various orchestration approaches provided by the 5G PPP projects and aiming at bridging an identified gap between the cloud computing orchestration solutions and the network services orchestration solutions, the multi-domain orchestration framework comes up with a novel and holistic approach for overall lifecycle of applications' design, development, deployment, and orchestration in a 5G ecosystem. Novel concepts include the separation of concerns among the orchestration of the developed applications and the underlying network services that support them, as well as the specification and management of application-aware network slices. In this top-down approach, application design and development lead to the instantiation of application aware-network slices, over which vertical industry applications can be optimally served.

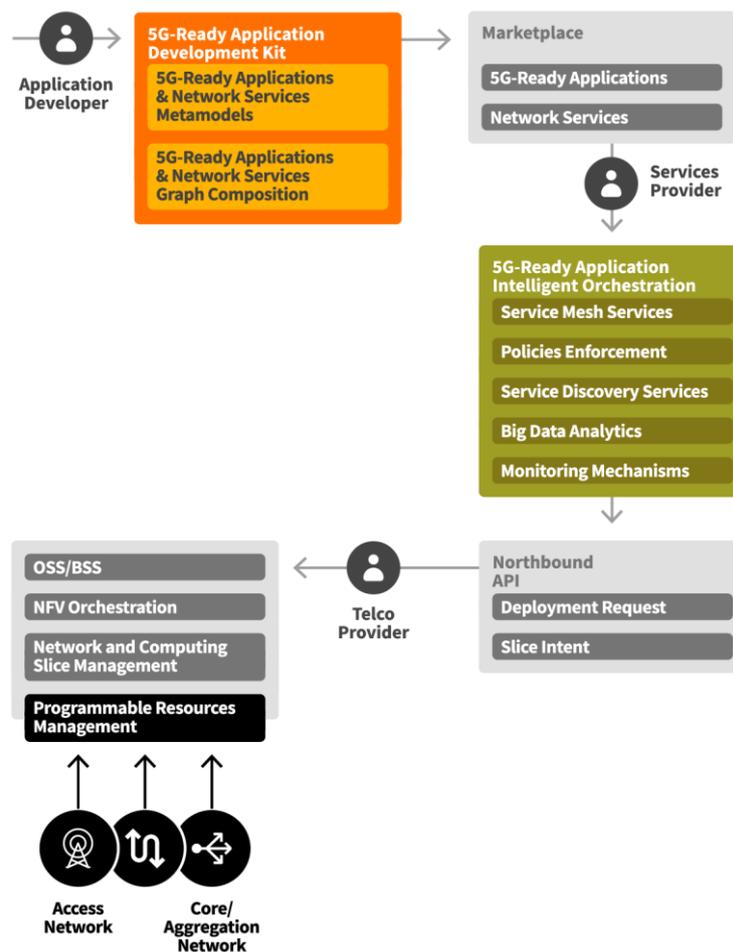


Figure 2-5: Framework for application-aware orchestration (adapted from [2-13])

The proposed framework (cf. Figure 2-5) is divided in three distinct layers, namely the Applications Layer, the Applications' Orchestration Layer and the Programmable Infrastructure Slicing and Management Layer. The Applications Layer is oriented to software developers, the 5G-ready Application Orchestration Layer is oriented to SPs and the 5G Infrastructure Slicing and Management Layer is oriented to NOPs, VISPs, and DCSP, cf. Figure 2-1. The Applications Layer takes into account the design and development of 5G-ready applications per industry

vertical, along with the specification of the associated networking requirements, which are tightly bound together with their respective applications' graph. The graph defines the business functions, as well as the service qualities of the individual application. The Applications' Orchestration Layer supports the dynamic deployment and service-aware adaptation of the applications, by using a set of optimisation schemes and intelligent algorithms that select the needed infrastructure resources from different domains. The Programmable Infrastructure Slicing and Management Layer is responsible for setting up and managing the application deployment and operation by means of an application-aware network slice. Network slice instantiation, activation, run-time management and orchestration, as well as monitoring are realized. According procedures can also be triggered by the Applications' Orchestration Layer via according APIs.

2.4.3 Service-specific extensions of ETSI NFV MANO

MANO systems enable an integrated and holistic approach towards NS and VNF management. While MANO processes can be standardized on a high level of abstraction, specifics of both the infrastructure and the service implementation need to be considered on the implementation level. For example, although a generic placement algorithm is able to instantiate all VNFs of, e.g., a content delivery network service at some location, the result might not be optimal. A specialized placement algorithm can bring in additional knowledge about the service and its components and, for example, place caches close to users.

As service developers have the best understanding of the MANO requirements and other peculiarities of their services, they are in the best position to add such service-specific knowledge to their services. To that end, a plug-in approach for NFV MANO systems allows the integration of service and function-specific managers into specific MANO processes. Service and function-specific management components are shipped together with their associated services as part of the service package. The managers are then integrated into the NFV MANO system, cf. Figure 2-6, and called as part of the generic, standardized processes of the MANO system on both the NFVO level and VNFM level.

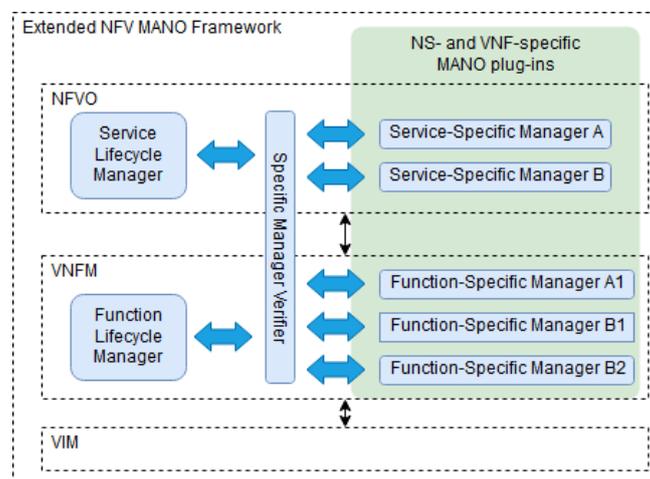


Figure 2-6: MANO plug-ins in the ETSI NFV MANO architecture

2.5 Programmable Networks

Network programmability has been one of the key paradigms in the development of 5G systems. Programmability ensures the flexible adaptation on different levels, including infrastructure, network functions, services and applications. The following sections provide examples for programmability of data plane, transport networks, and RAN functions, respectively.

2.5.1 Data plane programmability

The data plane landscape is composed by a wide range of heterogeneous resources, geographically grouped in three main tiers: radio access, edge nodes and central datacentres:

- Tier 1: edge area where radio access nodes are deployed;
- Tier 2: edge area with limited computing resources, corresponding to, e.g., street cabinets;
- Tier 3: central area with massive computing resources, corresponding to a datacentre.

All tiers provide features for programmability and flexible configuration by generating abstract views on resources of the underlying infrastructure. The solution consists of the utilization of the SDN paradigm to realize data plane configuration in a way that is agnostic to the underlying hardware infrastructure and fully integrated with management and orchestration plane.

The SDN architecture consists in the following layers, cf. Figure 2-7:

- WAN Resource Manager (SDN Application) is the functional element that triggers SDN control plane operations. It translates the abstracted view at orchestrator level in a network domain-specific view, ensuring that external link information contained at orchestrator level is translated in a suitable path between NFVI PoPs;
- Two types of SDN Controllers, one dedicated to the configuration of the network domain and the second dedicated to the configuration of the RAN domain; each controller is supported by according SDN agents located on the respective network elements;
- A data-plane consisting of Core NFVI, backhaul network, Edge NFVI, fronthaul network, WLAN Access Points and LTE small cells are the network elements and are considered as part of the infrastructure layer.

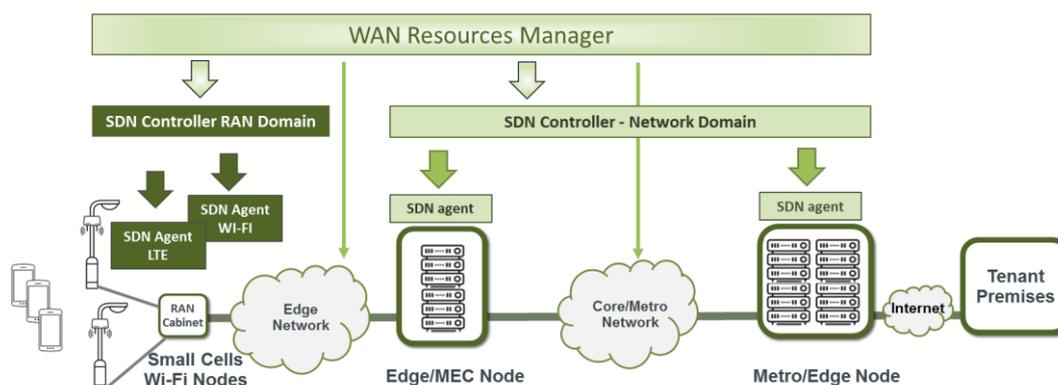


Figure 2-7: SDN architecture for data plane programmability [2-14]

Figure 2-8 depicts a concrete example for realizing the proposed approach for data plane programmability in a Cloud Enabled Small Cell (CESC) environment. A two-tier virtualised execution environment in the form of the Edge data centre allows the provision of SDN capabilities. On top, the CESC Manager (CESCM) triggers SDN control plane operations by translating the abstracted view at orchestrator level into network domain-specific views.

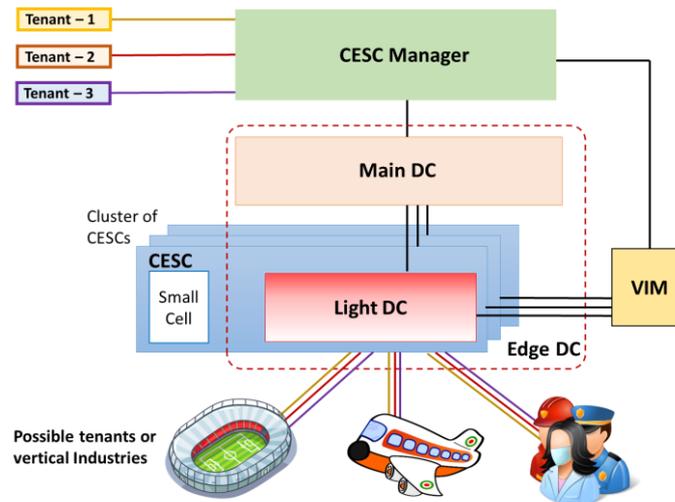


Figure 2-8: Proposed functional architecture [2-16]; DC stands for Data Centre

The SDN approach of decoupling control and data plane functions is suitable to make global decisions across several small cells, so called CESC clusters. The Virtualised Infrastructure Manager (VIM) controls the NFV Infrastructure (NFVI), which includes the computing, storage and networking resources of the edge data centres, and creates and controls the CESC clusters. Utilisation of small cells is partitioned into logically isolated slices, offered to different operators or tenants. The CESC Manager manages and orchestrates the logical cloud environment formed by the so-called “Light Data centre” and the small cell functions. Further, it coordinates and supervises the use of radio resources and service delivery. It controls the interactions between the infrastructure level and the network operators. For service assurance and fulfilment, CESC Manager encompasses telemetry and analytics functions for managing the overall network in an efficient and SLA-compliant manner. The CESC Manager functions will be built upon the services provided by the VIM for appropriately managing, monitoring and optimising the overall operation of the NFVI resources at the edge data centre. The NFV resources will be ultimately offered via a set of APIs that will allow the execution of network services over the distributed CESC.

2.5.2 Transport network programmability

Data plane programmability has been advocated as the perfect solution to manage the heterogeneity of 5G networks as well as to provide fast and easy network function deployment. In the transport network domain, solutions must adapt to the highly variable bandwidth requirements of future RANs, offering at the same time high levels of flexibility as well as resource and energy efficiency. The “DisAggregated RAN” [2-21] is a novel concept adopting the notion of “disaggregation” of HW and SW components across the wireless, optical and compute/storage domains. Apart from increased flexibility, disaggregation offers enhanced scalability, upgradability and sustainability potential. These features are particularly relevant when a continuously growing number of devices and services, as well as novel features, such as, the concept of flexible functional splits, need to be supported. “Resource disaggregation” decouples hardware and software components creating a common “pool of resources” that can be independently selected and allocated on demand. These components form the basic set of building blocks that can be independently combined to compose any infrastructure service. To exploit the concept of disaggregation in RAN environments, novel solutions must increase the density and power efficiency of the “pool of resources” and provide high bandwidth connectivity between them [2-22]. Such solutions will rely on i) hardware programmability: allowing HW repurposing to enable dynamic on demand sharing of resources, and ii) network softwarisation: enabling migration from the traditional closed networking model that focuses on network entities to an open reference platform that instantiates a variety of network functions.

Accordinging architectures take advantage of SDN to exploit the offered reconfigurability of high-performing switching hardware; and NFV's full programmability of network functions via software on commodity hardware platforms [2-23]. They adopt the concepts of transport network slicing and resource and service virtualisation across technology domains in order to develop a unified, programmable control and management framework [2-24] that can be used to coordinate the underlying heterogeneous technology domains and support end-to-end service provisioning across various infrastructure domains.

2.5.3 Network function programmability in RAN

The RAN architecture takes the baseline architecture, where the baseline architecture covers 5GPPP Phase 1 consensus and the 3GPP status from the publication time, i.e., the latest 3GPP Release specification on 5G RAN [2-3] [2-4], e.g., addition of Service Data Adaptation Protocol (SDAP) layer and F1 interface with CU-DU split. Here, the Controller Layer is envisioned for RAN [2-25], which provides means to introduce RAN control functions as specific application implementations. It is worth noting that such flexibility is already available for the CN thanks to the application functions (AFs) as part of the service-based architecture (SBA) [2-2]. A high-level illustration of the RAN architecture is given in Figure 2-9. Therein, the Controller layer is composed by cross-slice (XSC) and intra-slice controllers (ISC) along with the corresponding applications (APPs) running on the northbound interface (NBI). The control commands and interactions with the gNBs take place via the southbound interface (SoBI).

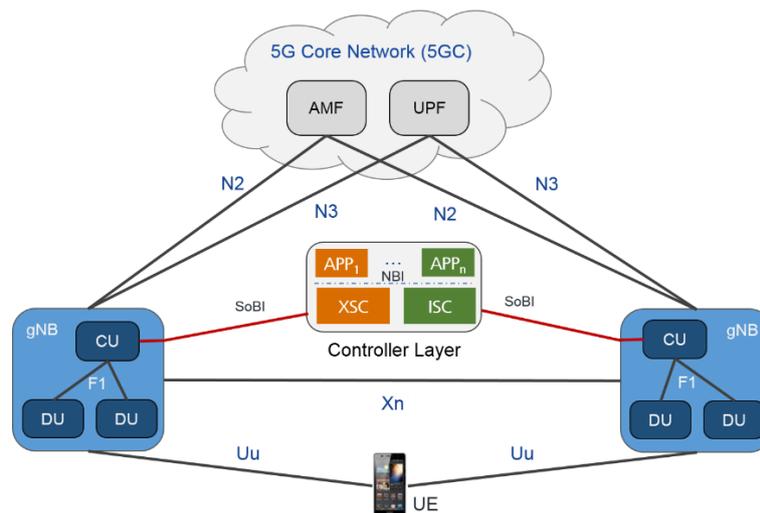


Figure 2-9: High-level RAN architecture with the controller layer providing RAN programmability

It is envisioned that the Controller layer communicates with the RAN NFs via the RAN Controller Agent (RCA), which is introduced in the CU to interface distributed and centralised NFs to the logically centralised controllers. In general, the RCA acts a middleware between controller and NFs with a local data-store capable to store most recent monitoring information from the NFs. In this regard, RCA can be considered as one of the common platform functions, cf. Figure 2-2. The amount of the data to be exposed to the Controller layer is thus controlled by the RCA. The SoBI is the unified interface between RCA and the controllers for monitoring and re-configuration of NFs. Each programmable NF in DU and CU supports interaction with RCA for exchanging control information with northbound applications deployed on top of the controllers. The RCA is interfacing the so-called RAN data analytics function (RAN-DAF), which is responsible for collecting monitoring information related to both UEs and RAN, such as Channel Quality Indicator (CQI), power level, path loss, radio link quality, radio resource usage, Modulation and Coding Scheme (MCS), Radio Link Control (RLC) buffer state information, etc. The RCA can

forward the information obtained from RAN-DAF to the controllers and further to northbound applications, such as, slow inter-slice RRM, slice-aware RAT selection, elastic resource control, etc. RCA also routes re-configuration information from controller to the respective NFs in the CU and DU.

2.6 Vertical-specific architecture extensions

The following subsections elaborate on service-specific extensions of the overall architecture shown in Figure 2-2. They demonstrate that the 5G system can be flexibly extended and customized to serve the requirements of vertical industries. This is illustrated using the examples energy utilities, vehicular communications, as well as media content production and delivery.

2.6.1 Extensions for energy utilities

The aim of the proposed extensions to the overall architecture as shown in Section 2.2 is to enable energy utilities in their transition towards more decentralized systems focusing on renewable energy and accelerate their digitalization. Relevant extensions include several VNFs offering SaaS and IaaS, Self-X functions as well as smart energy (application) VNFs. Since energy grids constitute a core part of critical infrastructures, guaranteed quality of service is crucial and self-optimization processes considering energy grid KPIs must be provided. Referring to Figure 2-2, the extensions focus on the Radio Access Network, the Management of Domain Resources and Functions and the E2E Service Creation.

Extensions in the **Radio Access Network** realise new methods for IoT device identification and optimization of data routing for small and very small devices. On the Resources and Function Level, this comprises application specific VNFs, deployed at the edge. These VNFs focus on i) the extensive monitoring of the energy grid and networking infrastructure, ii) the digitization of the existing control of the energy grids, iii) the decoupling of the smart grid assets from the physical devices by means of employing so-called digital twin technologies, iv) the introduction of blockchain technologies towards storing critical data in an unambiguous, traceable manner, v) the acceleration of infrastructures maintenance- and security-oriented media, and vi) the enabling of high-accuracy mobility management services allowing for better management of next-generation devices such as drone swarms for automated inspection.

Extensions in the **Management of Domain Resources and Functions** enable the service-aware configuration and orchestration of specific resources and functions. On the Network Level, such adaptations can be used to create isolated end-to-end network slices on the same infrastructure for simultaneous use by heterogeneous services. Indeed, depending on the operational environment, the energy utility vertical uses all three 5G flavours: eMBB (e.g. drones for remote infrastructure inspection), mMTC (e.g. connecting 5G-ready advanced smart metering infrastructure deployments), and URLLC (e.g. connecting scalable installations of phasor measurement units). Additionally, to better coordinate different resource categories, analytics-based optimization mechanisms, which are controlled by a utility-based policy, govern the behaviour of network services and also consider application-level metrics, e.g., energy grid-related KPIs. To this end, two new interfaces are introduced linking the Analytics component with the Service Operations component (Operations-Analytics, Os-An interface) and with the Domain Management component (Analytics-Management, An-Ma interface), respectively.

Finally, in the **E2E Service Creation**, various multi-tenant applications and specific ‘Smart Energy as a Service’ applications are deployed. Indeed, 5G-enabled energy grids would enable ‘killer applications’ such as advanced metering infrastructure as a service, predictive maintenance as a service, as well as dispatchable demand response as a service, which have the potential of revolutionising the operations workflow of energy utilities.

2.6.2 Extensions for vehicular communications

Vehicular communications simultaneously involve multiple use cases, traffic types, and communication paths. In fact, in addition to transmissions routed through the core network towards remote servers, links between vehicular UEs in proximity may involve the PC5 link for direct Vehicle-to-Vehicle communications, whereas local breakout can be applied for network-assisted links routed through the edge network. In this latter case, the base station or road side unit can locally relay the messages to the UEs in proximity, and/or route them to UEs attached to neighbouring base stations via short routing paths passing through the edge data centre.

Automotive applications include a wide set of services, offered by different providers, each imposing a specific set of requirements. For this reason, vehicular communications rely on the network slicing feature, where the lifecycle management of each slice is tailored to support the related service, cf. Figure 2-10. This collection is composed by slices belonging to the standard types already defined by 3GPP, notably eMBB and URLLC, thus exploiting the flexibility provided by the standardised slice types.

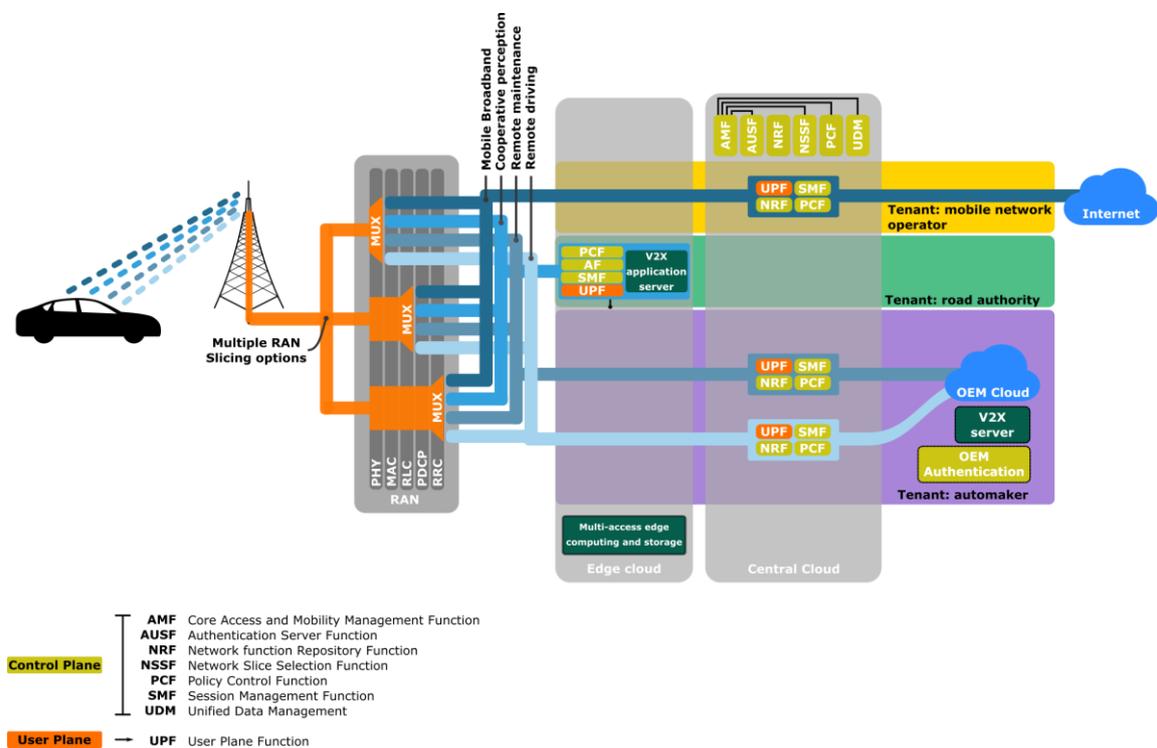


Figure 2-10: Example of slice-based architecture for vehicular applications

For the automotive vertical, it is important that network functions can be deployed both in the edge and central cloud, according to the requirements they are designed to serve. The edge cloud hosts NFs which need to be allocated in proximity of the UEs, potentially including additional features such as Multi-access Edge Computing (MEC) and storage facilities. The central cloud, on the other hand, contains the slice-specific network functions for use cases requiring connectivity with a remote public network.

The concept of multi-tenancy is leveraged in vehicular applications, wherein the tenant is the company, vertical, or service provider offering the services supported by one slice, or one set of slices. Examples of tenants for automotive applications are mobile network operators, road operators, and automakers.

Road authorities may provide Cooperative Intelligent Transportation System (C-ITS) services like hazard warning, in-vehicle signage, and in general cooperative perception and cooperative

manoeuvre services. These services involve information that is both strictly time-sensitive and location-sensitive: messages are in fact transmitted and received by vehicles to spread and acquire safety-critical information about the instantaneous traffic conditions in their surroundings. These services hence require a low latency slice with high reliability, providing timely reception of these messages. For this reason, network resources are foreseen to be mostly allocated in the edge cloud, as close as possible to the road users. Alternatively, sufficient transport network resources towards the central cloud must be allocated for the slice.

Automakers may offer different classes of services to their clients, such as remote maintenance and tele-operated driving. Both require connectivity between the vehicle and the automaker's cloud, although each with completely different service level requirements. In the former case, the machine-type communications, which could be delivered via an eMBB slice, is used to retrieve data from the on-board sensor to plan ahead the maintenance of large vehicular fleets. In contrast, remote driving requires low latency, high data rate, and high reliability in the uplink to provide a real-time video flow and instantaneous sensor data to the remote driver. Similarly, the downlink needs to deliver the driving commands to the vehicle. While both services mostly rely on network functions running in the operator's central cloud, they have completely different degrees of redundancy. Furthermore, the automaker can implement further authentication functions beyond those offered by the network, as well as hosting Vehicle-to-Everything (V2X) application servers in their premises. In any case, network slices can leverage dedicated bearers to provide tailored QoS to specific applications/flows also within the same slice. Distinct applications are assigned to specific bearers (in the RAN), and flows (in the core network), which are treated with a different level of priority.

2.6.3 Extensions for enhanced content delivery

Within the broad variety of services 5G networks target, there are use cases that requires content delivery to a group of end devices using broadband connectivity over mobile and converged networks. Example of such use cases are live video streaming, mission critical communication, information dissemination in IoT and V2X domains. Live video streaming has been growing for many years due to improvements in network performance which made HTTP streaming a viable solution for delivering live media. Nonetheless, live video streaming creates several challenges when delivered over unicast connections. In particular, demand for network throughput may experience significant, yet short-term, peaks, forcing network operators to overprovision their networks to deal with such peak demands [2-26].

A novel framework allows content delivery over mixed network types comprising fixed and mobile, as well as unicast, multicast and broadcast connection types, cf. Figure 2-11 [2-27]. The framework introduces two logical functions to handle multicast transport in an end-to-end manner: 'Function X' on the multicast server end and 'Function Y' on the multicast termination end. Function X handles the encapsulation of (unicast) data into multicast. The placement of Function X depends on many aspects including capabilities of underlying network fabric, cache placement strategies, etc. In this simplified representation of the framework, it is recognized that Function X may be located either outside of converged network (Function X.1) or within converged network (Function X.2). Generally, Function Y would exist in the Home Gateway, or possibly partly in the UE, and would accept the input from the upstream network in both unicast and multicast form, and present unicast to the downstream clients.

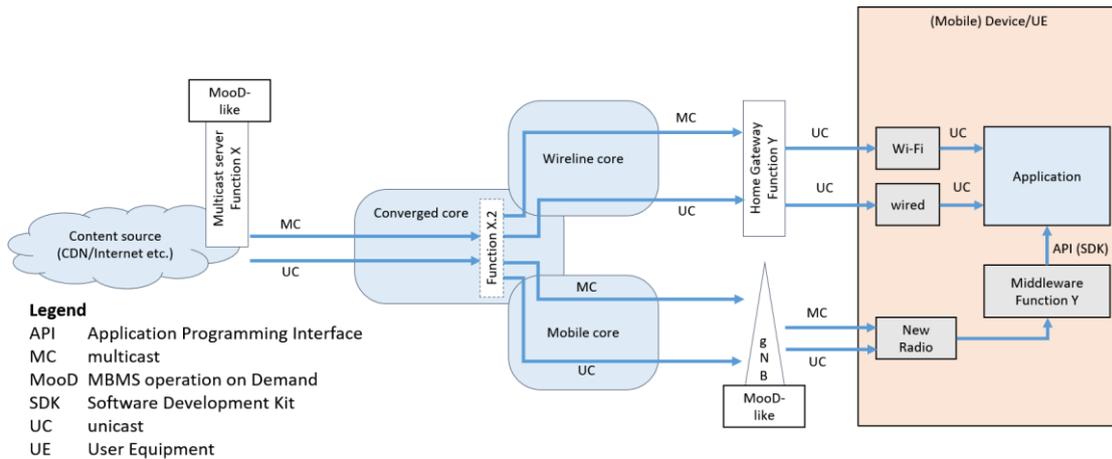


Figure 2-11: Framework for efficient content delivery over 5G networks [2-27].

2.6.4 Extensions for media production and delivery

For enhanced support of media production and delivery via 5G networks, Function-as-a-Service (FaaS) technology, (such as OpenWhisk), should become integral part of cloud services [2-11]. FaaS addresses use cases that happen spontaneously and require immediate setup of an elastic communication service. Such an approach aims at overcoming today’s limitations posed on traditional broadcast productions by implementing orchestrated mobile content contribution, remote and smart media production, and low-latency and high-bandwidth media distribution (e.g., streaming) over 5G networks. Figure 2-12 shows VIM extensions for supporting FaaS that are compatible with the ETSI MANO reference architecture. The “FaaS-VIM” supports the standard Vnfm-Vi, Or-Vi, and Nf-Vi interfaces on the northbound. Southbound interfaces of the FaaS-VIM are specific to the FaaS framework being used for implementation. The FaaS framework can be deployed on some underlying PaaS that, in turn, can use IaaS virtualisation technology or directly run on the bare metal machines. The rationale of the architecture is to allow compatibility with the ETSI NFV standard without tightly coupling the FaaS paradigm to some specific implementation or deployment options.

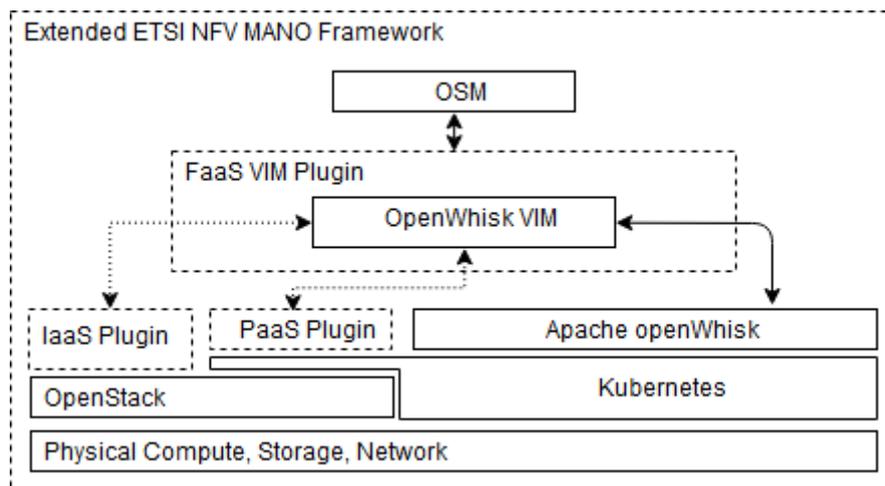


Figure 2-12: Extending ETSI MANO Reference Architecture with FaaS

A typical workflow looks as follows: a journalist covers an important event and streams a live signal via smartphone back to the broadcaster’s facilities. The faces depicted in the video stream trigger the Face Recognition Engine, which instantiates the FaaS capabilities automatically. As soon as it is up and running, the Face Recognition Engine starts tagging the people in the video

frames and sorting the material under the right tags in the broadcaster's archive. If the recognised content in the frames is part of a current story, another FaaS could notify an editor.

2.7 References

- [2-1] 3rd Generation Partnership Project (3GPP), Technical Report TR 28.801 "Telecommunication management; Study on management and orchestration of network slicing for next generation network," 2017
- [2-2] 3rd Generation Partnership Project (3GPP), Technical Specification TS 23.501 "System architecture for the 5G System (5GS)," 2019
- [2-3] 3rd Generation Partnership Project (3GPP), Technical Specification TS 38.300 "NR; Overall description; Stage-2," 2019
- [2-4] 3rd Generation Partnership Project (3GPP), Technical Specification TS 38.401 "NG-RAN; Architecture description," 2019
- [2-5] Next Generation Mobile Networks Alliance (NGMN), "Description of Network Slicing Concept by NGMN Alliance, NGMN 5G P1 Requirements & Architecture Work Stream End-to-End Architecture," 2016
- [2-6] 5G-TRANSFORMER, Deliverable D1.1 "Report on vertical requirements and use cases," 2018, retrieved from <http://5g-transformer.eu/>
- [2-7] SliceNet Deliverable D2.2 "Overall Architecture and Interfaces Definition", https://doi.org/10.18153/SLIC-761913-D2_2
- [2-8] 5G-VINNI Deliverable D5.1 "Ecosystem analysis and specification of B&E KPIs", <https://doi.org/10.5281/zenodo.3345665>
- [2-9] 5G-TRANSFORMER, Deliverable D1.3 "5G-TRANSFORMER refined architecture," 2018, retrieved from <http://5g-transformer.eu/>
- [2-10] 5GTANGO, "5G Development and validation platform for global industry-specific network services and Apps," 2019, retrieved from <http://5gtango.eu/>
- [2-11] 5G-MEDIA, Deliverable D3.2 "Specification of the 5G-MEDIA Serverless Computing Framework," 2018, retrieved from <http://www.5gmedia.eu/>
- [2-12] Next Generation Mobile Networks Alliance (NGMN), "5G End-to-End Architecture Framework," 2018.
- [2-13] MATILDA, Deliverable D1.1 "MATILDA Reference Architecture, Conceptualization and Use Cases," 2017, retrieved from <http://www.matilda-5g.eu/>
- [2-14] 5GCity, "A Distributed Cloud & Radio Platform For 5G Neural Hosts," 2019, retrieved from <http://www.5gcity.eu/>
- [2-15] MATILDA, "A Holistic, Innovative Framework for Design, Development and Orchestration of 5G-ready Applications and Network Services over Sliced Programmable Infrastructure," 2019, retrieved from <http://matilda-5g.eu/>
- [2-16] 5G ESSENCE, Deliverable D2.2 "Overall System Architecture and Specifications," 2018, retrieved from <http://5g-essence-h2020.eu/>
- [2-17] ETSI NFV ISG, "NFV Proof of Concept," 2019, retrieved from <https://www.etsi.org/technologies/695-network-functions-virtualisation>
- [2-18] TOSCA, "TOSCA Simple Profile in YAML," V1.2, 2017, retrieved from <https://www.etsi.org/technologies/695-network-functions-virtualisation>
- [2-19] Internet Engineering Task Force, "YANG – A Data Modelling Language for the Network Configuration Protocol," 2010, retrieved from <https://tools.ietf.org/html/rfc6020>
- [2-20] Kapassa, E., "SLAs in 5G: A Complete Framework Facilitating VNF- and NS- Tailored SLAs Management," Int'l Conference on Advanced Information Networking and Applications Workshop, pp. 469-474, 2018.
- [2-21] 5G-PICTURE, Deliverable D2.2 "System architecture and preliminary evaluations," 2018, retrieved from <https://www.5g-picture-project.eu/>

- [2-22] 5G-PICTURE, Deliverable D3.2 “Intermediate report on Data Plane Programmability and infrastructure components,” 2018, retrieved from <https://www.5g-picture-project.eu/>
- [2-23] 5G-PICTURE, Deliverable D4.2 “Complete design and initial evaluation of developed functions,” 2018, retrieved from <https://www.5g-picture-project.eu/>
- [2-24] 5G-PICTURE, Deliverable D5.1 “Relationships between Orchestrators, Controllers, slicing systems,” 2018, retrieved from <https://www.5g-picture-project.eu/>
- [2-25] 5G-MoNArch, Deliverable D2.3 “Final overall architecture,” 2018, retrieved from <https://5g-monarch.eu/>.
- [2-26] 5G-Xcast, Deliverable D5.1 “Content Delivery Vision,” 2017, retrieved from <http://5g-xcast.eu/>
- [2-27] 5G-Xcast, Deliverable D5.2 “Key Technologies for the Content Distribution Network,” 2018, retrieved from <http://5g-xcast.eu/>

3 Radio & Edge Architecture

3.1 Overall RAN Architecture

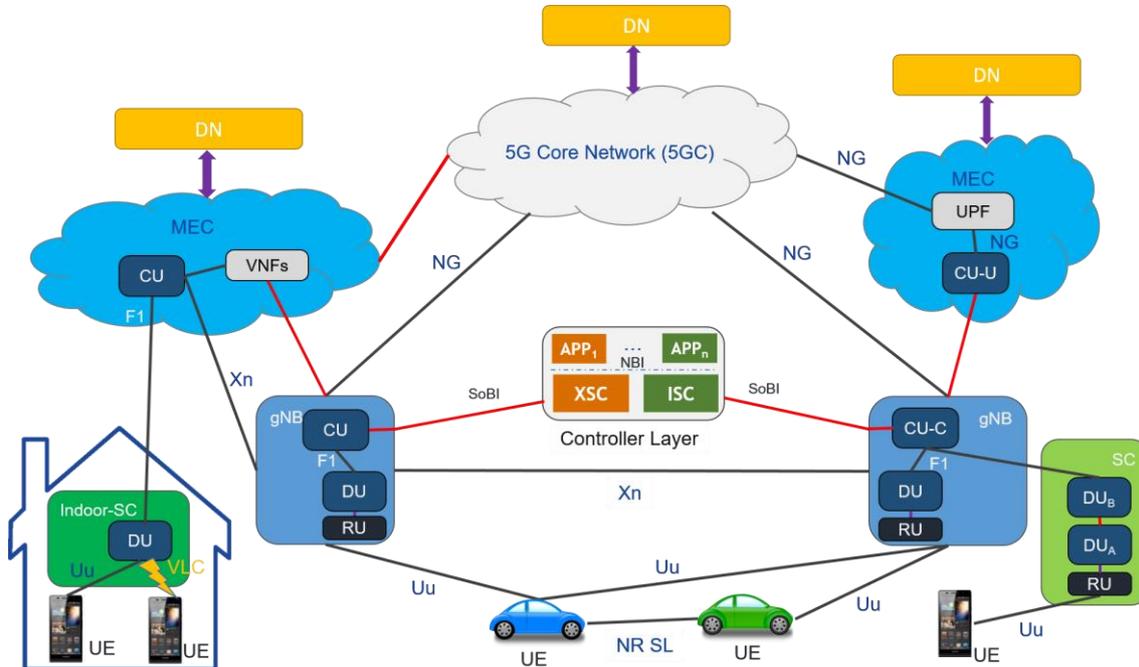


Figure 3-1: Overall RAN Architecture

The overall Radio Access Network (RAN) architecture, depicted in Figure 3-1, is based on a baseline architecture including 5GPPP Phase 1 consensus and the latest 3GPP Release specifications on NG-RAN [3-1][3-2], comprising the addition of the Service Data Adaptation Protocol (SDAP) layer, and of the F1 interface with Centralized Unit – Distributed Unit (CU-DU) split. The innovative architecture is capable of providing Small Cell (SC) coverage to multiple operators “as-a-Service”, enriched with a two-tier architecture including a first distributed tier for providing low latency services, and a second centralized tier for providing high processing power for compute-intensive network applications. The versatility of the architecture is further enhanced by high-performance virtualization techniques for data isolation, latency reduction and resource efficiency, and by orchestrating lightweight virtual resources enabling efficient Virtualized Network Function (VNF) placement and live migration. In particular, the suggested solution envisages to virtualize and to partition small cell capacity, while at the same time it aims to support enhanced edge cloud services, by enriching the network infrastructure with an edge cloud. It is worth noting that in addition to the terrestrial communications, the radio access can also be realized by non-terrestrial technologies, as described in Section 3.2.8.

The CU may be further split into Control Plane (CP) part, referred to as CU-C or CU-CP, and the User Plane (UP) part, referred to as CU-U or CU-UP. Such split enables the implementation of CU-C and CU-U in different locations, as depicted in Figure 3-1. Another additional split option is the lower layer split, which can be applied to a DU. It is worth noting that a DU can operate as a small cell. The RAN architecture enhances the baseline architecture by functional models emerging from the 5GPPP innovations as outlined in the following sections. One such extension is the controller layer, which enables RAN programmability (see Section 2.5.1), in terms of RAN control functions, as specific application (APP) implementations. The APPs can run on the North-Bound Interface (NBI) over the cross-Slice Controller (XSC) and Intra-Slice Controller (ISC), and the communication with the RAN can be maintained over the South-Bound Interface (SoBI).

It can be envisioned that such APPs provide slow-scale control functionality and can support the RAN control functions, such as the Radio Resource Management (RRM). To satisfy the most demanding use cases, including safety-critical vehicular applications, the cooperation of multiple links is envisioned; furthermore, local end-to-end paths are introduced to minimize the latency between vehicles and road users located in respective proximity.

The use of Network Function Virtualization (NFV) technology enables the deployment of multiple small cells (such as visible light communication gNBs in buildings) with no cost of signalling with the 5G Core. As depicted in Figure 3-1, multiple small cells can be seen as one small cell by deploying a VNF in the cloud as a form of Multi-access Edge Computing (MEC). The other possible deployment scenarios for small cells are either to have all small cells connected directly to the 5G Core using NG interface, thereby seen and managed by the 5G Core, or to adopt dual connectivity mode. The use of VNFs reduces the signalling significantly compared to the other possible deployment scenarios, as described in Section 3.3.2.

3.1.1 Centralized-Distributed-Radio Units split and Control-User Planes separation

In 5G PPP Phase 2 and Phase 3 projects, a number of key technologies defined in 3GPP are taken as the baseline, and enhanced with additional specific extensions to meet the requirements of the individual projects. Most implementations include CU-DU split, with some going further to also include exposure of Common Public Radio Interface (CPRI) creating a split in the radio equipment between a Remote Unit (RU), Distributed Unit (DU) and Centralized Unit (CU), such as documented in [3-3].

In [3-3], possible options for decomposition of the RAN environment are studied, resulting in the identification of eight options, illustrated in Figure 3-2.

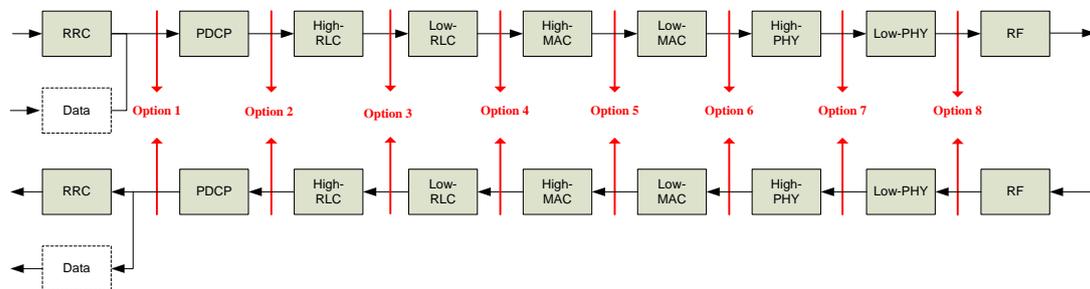


Figure 3-2: RAN decomposition options [3-3]

Option 8 from Figure 3-2 is the exposure of the CPRI interface, while Option 7 is referred to as enhanced CPRI (eCPRI). Both amount to the separation of the RU from the Base Band Unit (BBU). Of the remaining options, only Option 2 has resulted in significant further work. Option 2 consists in the separation of a distributed unit and a centralized unit, with the F1 Reference point defined to connect the CU and DU.

Furthermore, work was conducted to separate the CU into its UP and CP components, as documented in [3-5] and illustrated in Figure 3-3, which highlights the decomposition of the F1 interface into control and user plane parts, and the exposure of the E1 reference point between the control plane and user plane functions of the CU.

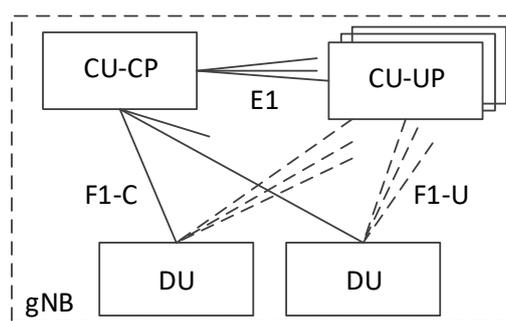


Figure 3-3: CP/UP split of gNB [3-5]

3.2 Protocol Extensions for Vertical Support

3.2.1 RAN Part of Network Slicing

3.2.1.1 Fundamental Slicing Support in RAN

A fundamental support for network slicing is provided in the RAN by the 3GPP specifications: 3GPP Release 15 for Next Generation-RAN (NG-RAN) was frozen in June 2018, and a so-called late drop of Release 15, including further architecture options, was initially planned to be frozen by the end of 2018 [3-6], then extended to the end of March 2019 [3-7]. This specification comprises slicing awareness in RAN via Network Slice Selection Assistance Information (NSSAI), including one or more Single NSSAIs (S-NSSAIs), which allow to uniquely identify network slices [3-1]. While the fundamental slicing support is achieved by Release 15, including granularity of slice awareness and network slice selection, various enhancements and optimization can be considered for future releases, implying for example specification-relevant signalling changes, and implementation-dependent algorithms, for instance related to the management of shared resources between slices.

Network slicing offers unprecedented degrees of flexibility, since Network Functions (NFs) can be tailored according to the diverse Quality of Service (QoS) requirements of different slice tenants. For instance, these may only include performance differentiation in terms of latency and data rate, or be extended by further Service Level Agreement (SLA) requirements, such as the number of connections for a given time and location. Therefore, slice tenant requirements can be supported by different network slicing implementation variants [3-8]. In some of these variants, the whole RAN protocol stack can be shared by network slices whose SLA differentiation can be performed with QoS enforcement. In particular, in line with the latest 5G Release 15 specifications, and as shown in Figure 3-4, for a network slice instance one or more Protocol Data Unit (PDU) sessions can be established by the 5G Core (5GC); a PDU session belongs to one and only one specific network slice instance [3-9]. In Figure 3-4 the user planes of two network slice instances are shown, namely the enhanced Mobile Broadband (eMBB) and Low-Latency eMBB (LL-eMBB). These comprise the User Plane Functions (UPFs), the interface between the UPF and the RAN (referred to as NG-U, or N3 interface), the UP processing in the RAN, and the User Equipment (UE) connected via the Uu radio interface. The RAN maps packets belonging to different PDU sessions to different Data Radio Bearers (DRBs), where within a PDU session there can be one or more QoS flows [3-1]. On this basis, the RAN treatment of different network slices can be in terms of RRM schemes, performed based on the QoS profiles of QoS flows mapped onto the respective DRBs. QoS profiles include performance characteristics such as the Packet Delay Budget (PDB), the Packet Error Rate (PER), and the Allocation and Retention Priority (ARP).

Inter-slice (or multi-slice) Resource Management (RM) is thus very important for improving the system efficiency, especially on shared infrastructure resources, as it represents a means for cross-slice optimization [3-8] [3-10]. The inter-slice RM thus factors in the slice SLAs, for instance by adapting the instantaneous radio resource allocation. In addition to the slice-adaptive radio resource allocation, slice awareness can be extended to the so-called hard network resources, namely the wireless access nodes, particularly in self-backhauled dynamic small cells. That is, the slice support may not only include the conventional radio resources like time and frequency resources, but it can also include the adaptation of the network topology, considering the dynamic small cells available in a certain region; this is referred to as the extended notion of a resource. Accordingly, the slice-adaptive resource control shall consider the changing radio topology including different access node types, such as micro-cells, pico-cells, relays, and Vehicular Nomadic Nodes (VNNs).

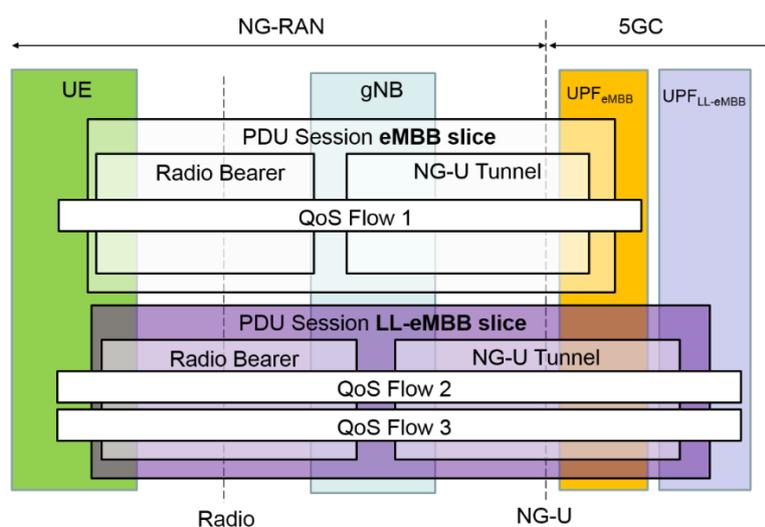


Figure 3-4: Network Slice support in the 5GS.

3.2.1.2 Wi-Fi virtualization and slicing

5G integrates different types of radio technologies, such as evolution of LTE, the 5G New Radio, and Wi-Fi-based technologies. Network slicing is one of the core mechanisms in 5G networks, requiring the instantiation of multiple virtual networks over a single, shared physical infrastructure. In wireless network mediums, virtualization is required to enable network slicing, and can be performed in different ways. A simple way to virtualize wireless interfaces is to share a wireless interface among a set of tenants. For example, in LTE, a Public Land Mobile Network (PLMN) ID can be instantiated for each tenant on the same carrier, to differentiate between the tenants.

One possible implementation of Wi-Fi virtualization is developed in the framework of the Linux operating system and it is achieved instantiating virtual wireless interfaces running in the “user space” that leverage the mac80211 module, running in the “kernel space”. The hardware drivers bridge the mac80211 kernel module with the physical Wi-Fi Network Interface Card (NIC), as shown in Figure 3-5. Wi-Fi supports a series of different types of virtual interfaces (vifs), including virtual access points, and virtual mesh interfaces. In practice, each virtual access point has its own Service Set Identifier (SSID) that is announced with dedicated beacons, as a physical access point would do. This type of virtualization allows, for example, to instantiate dedicated SSIDs for particular tenants or services, along with specific settings for critical concepts as security (WPE/WPA/WPA2, etc.).

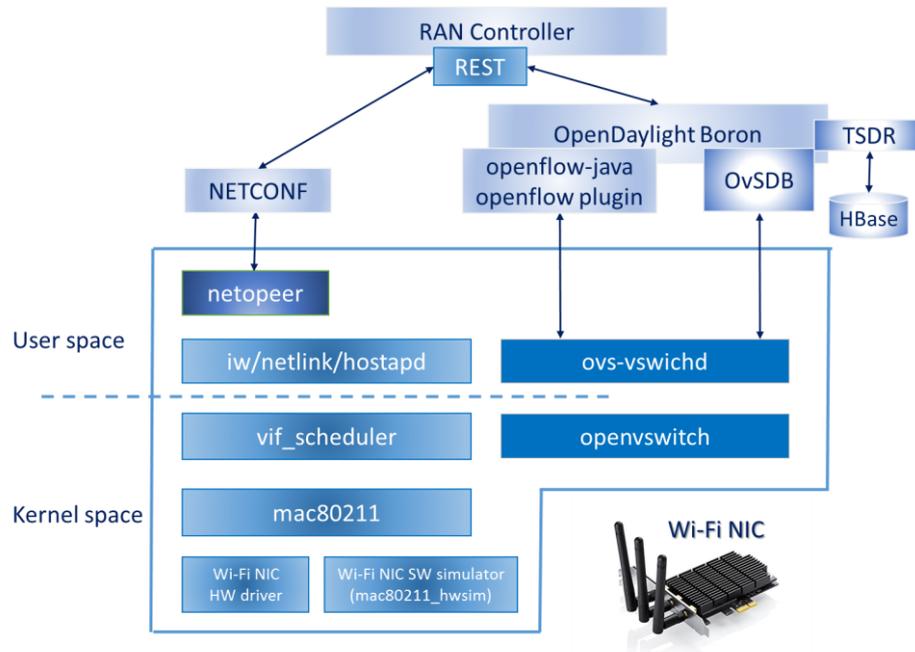


Figure 3-5: Wi-Fi virtualization and slicing

In this implementation example, the RAN slicing is realized leveraging two additional software modules: the virtual interface (vif) scheduler and the netopeer agent [3-11]. The vif scheduler is responsible for applying isolation in form of airtime slicing of the available radio resources, whereas the netopeer agent hosts a Network Configuration Protocol (NETCONF) server [3-12] that is used to set up and configure the virtual interfaces on top of the physical interfaces.

In wireless communications, while some upper data rate limits can be determined from the hardware specification, the actual available data rate of a wireless link can vary significantly. In particular, each User Equipment (UE) connected to a Wi-Fi access point can have a different nominal data rate from other users attached at the same time, due to the position of the equipment, obstacles, or even mobility; furthermore, different user equipment may not support all data rates offered by an access point. Another consideration for availability of wireless resources is that as more users get connected to an access point, the more the actual data rate decreases due to the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) access scheme implemented in Wi-Fi. Thus, it is not possible to guarantee specific data rates to a tenant as part of their network slice. In this case, a Wi-Fi RAN slice is therefore defined as the assignment of a ratio of the actually available radio resources, in terms of airtime. The airtime is defined to be the real time the transmission of a packet occupies the radio medium. This type of slicing is managed by the vif scheduler, which is composed of two parts:

- the local scheduler, which is an agent software running in the wireless nodes. This software is implemented as a dynamically loadable kernel module that sits on top of the mac80211 module. The scheduler can be configured to apply specific airtime ratios for any underlying virtual access points in the downstream traffic.
- the global scheduler, which is a software module that forms part of the SDN controller. The global scheduler is responsible for configuring the airtime ratios of the local schedulers and for monitoring them in order to detect whether the ratios are correctly applied.

Once a physical interface has been virtualized with one or multiple virtual interfaces running on top of it, it is possible to use SDN software elements to generate network slices. The basic mechanism consists in adding the virtual interfaces to the virtual software switches like Open

virtual Switch (OvS) [3-13], adding them effectively to the data plane of the SDN-based solution. The RAN controller offer Representational State Transfer (REST) APIs towards the NETCONF module and the OpenDaylight (ODL) SDN controller [3-13]; moreover the NETCONF module uses the NETCONF protocol to manage the RAN elements, whereas the SDN controller handles the Wi-Fi nodes using OpenFlow [3-15] and Open vSwitch Database (OvSDB) [3-16].

3.2.2 RAN Analytics

The RAN Data Analytics Function (RAN-DAF) is proposed with the motivation that real-time analytics are required for improving RAN NFs, like radio resource management. Since the RAN needs to enact fast decisions, the analytics based on the processing of real-time measurements shall stay local for optimizing performance dynamically. Moreover, from the business standpoint, different stakeholder may be involved among RAN, Core Network (CN), and Management. Hence, the storage and analysis of radio-related measurement may be restricted to be abstracted to CN or Operations Administration and Maintenance (OAM). An example deployment of such functionality is shown in [3-17], where more complex RAN deployments with CU-DU splits, better motivate for such functionality. Different options for performing RAN analytics may be examined: RAN-DAF may either be a control functionality in the RAN, or a management / Self-Organizing Network (SON) functionality. With the proposed Service-Based Architecture (SBA) being envisioned for both control and management functionalities, both implementations of RAN-DAF will rely on the inter-domain message bus interface.

Given the types of analytics and the proposed architecture enhancements (further elaborated in [3-17]), Table 3-1 provides some exemplary functionalities that can be defined and configured based on different slice requirements and network conditions.

Table 3-1: Analytics Functionality placement and classification [3-17]

	Parameter	Type	Time-scale
A. UE-related parameters	Mobility	Prescriptive Analytics	Real-time
	Interference level	Predictive Analytics, Prescriptive Analytics	Real-time
B. Network-related parameters	Radio Resource Situation (conditions, usage, availability)	Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics	Real-time / non-real time
	Backhaul Conditions / Availability (e.g. Base Station neighbourhood change)	Descriptive Analytics, Predictive Analytics	Non-real time
	Network QoS	Diagnostic Analytics, Predictive Analytics	Real-time / non-real time
	NW assistance (for V2X)	Predictive Analytics	Real-time / non-real time

3.2.3 Virtualized Small Cell (Cloud-enabled small cell) at the Light Data Centre

Current virtualization technologies considered in the proposed architecture allow mobile core functions to be deployed close to the mobile edge, thus enabling service delivery in a more effective way in proximity of final users [3-18]. For this reason, a two-tier virtualized execution environment is envisioned, materialized in the form of the edge data centre, which allows the provision of MEC capabilities to the mobile operators, enhancing the user experience and the agility in the service delivery.

The first tier is the Light data centre, hosted inside the Cloud Enabled Small Cells (CESCs), which supports the execution of the VNFs making up the Small Cell access. The Light data centre is envisioned to host network functions supporting traffic interception, GPRS Tunnelling Protocol (GTP) encapsulation/decapsulation, and some distributed RRM and SON. VNFs requiring low processing power, such as Deep Packet Inspection (DPI) and Machine-to-Machine (M2M) Gateway, could also be hosted in it. The connection between the small cell Physical Network Functions (PNFs) and the small cell VNFs can be realized through the network Functional Application Platform Interface (nFAPI). Finally, backhaul and fronthaul transmission resources will be part of the CESC, allowing for the required connectivity.

The second cloud tier, referred to as the Main data centre, is meant to host more computation-intensive tasks and processes that need to be centralized in order to have a global view of the underlying infrastructure. This encompasses the centralized Software-Defined Radio Access Network (cSD-RAN) controller, which is implemented as a VNF running in the Main data centre and makes control plane decisions for all the radio elements in the geographical area of the CESC cluster, including the centralized Radio Resource Management (cRRM) over the entire CESC cluster. Other potential VNFs that could be hosted by the Main data centre include security applications, traffic engineering, mobility management, and in general any additional network End-to-End (E2E) service that can be deployed and managed on the virtual networks, effectively and on demand.

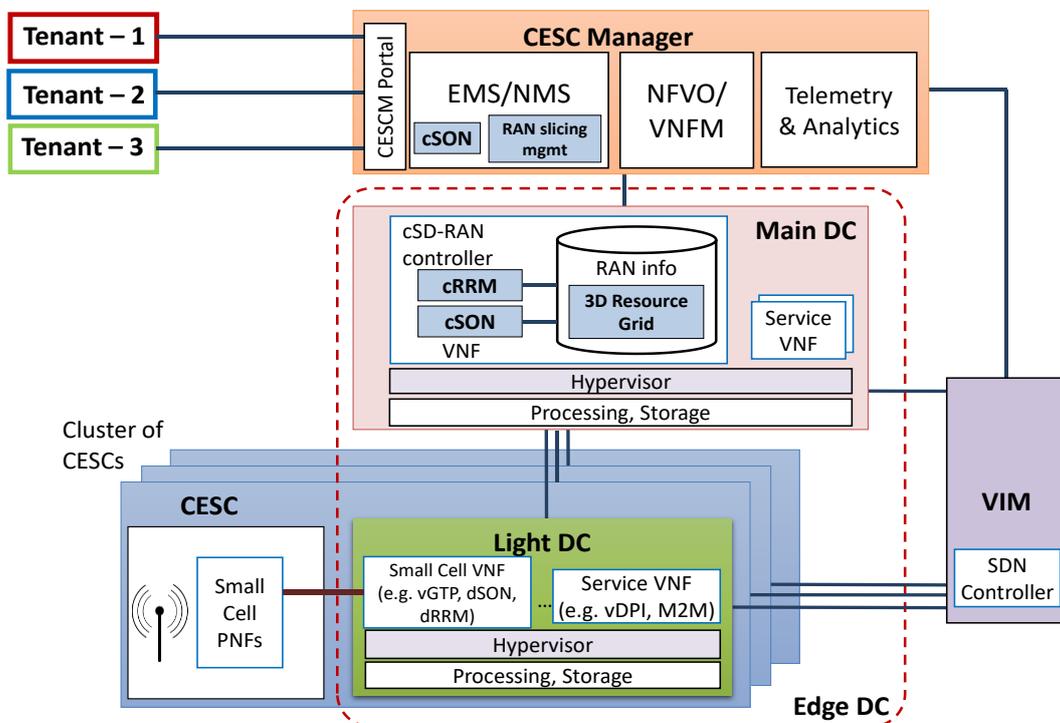


Figure 3-6: example of 5G High level architecture; DC stands for Data Centre

The Cloud Enabled Small Cell. In our scope, a CESC consists of a multi-Radio Access Technology (RAT) 5G small cell, with its standard backhaul interface, standard management connection [3-19], and with necessary modifications to the data model [3-20] to allow Multi-Operator Core Network (MOCN) radio resource sharing. The CESC is composed by a physical small cell unit attached to an execution platform based on one kind of microprocessor architecture, such as x86, ARMv8, or MIPS64. Edge cloud computing and networking are realized through the sharing of computation, storage and network resources of those micro servers present in each CESC, and form the Light data centre for implementing different features/capabilities of the Small Cells. Therefore, the CESC becomes a neutral host for network operators or virtual network operators that want to share IT and network resources at the edge of the mobile network.

The CESC is meant to accommodate multiple operators (tenants) by design, offering Platform as a Service, capable of providing the deployed physical infrastructure shared among multiple network operators. Different VNFs can be hosted in the CESC environment for different tenants. This also provides the support for mobile edge computing applications deployed for each tenant that, operating very near to the end users, may significantly reduce the service delivery time and deliver composite services in an automated manner. Moreover, the CESC is the termination point of the GTP-User Plane (GTP-U) tunneling protocol, which encapsulates user IP packets from the core network entities, such as the Evolved Packet Core (EPC) Serving Gateway (SGW) in LTE, destined to the UE and vice versa.

The CESC exposes different views of the network resources: per-tenant small cell view, and physical small cell substrate, which is managed by the network operator, decoupling the management of the virtual small cells from the platform itself. In the CESC such fan-out is done at the Light data centre, rather than providing multiple S1 (or Iu-h interface) connections from the physical small cell to different operators' EPC network elements such as Mobility Management Entity (MME) and SGW. Furthermore, the CESC is the termination of multiple S1 interfaces connecting the CESC to multiple MME/SGW entities as in S1-Flex. The interconnection of many CESC forms a 'cluster' which can facilitate access to a broader geographical area with one or more operators (including virtual ones), extending the range of their provided services while maintaining the required agility to be able to provide these extensions on demand.

The Edge data centre encompassing Main data centre and Light D data centre. The Edge data centre envisages combining the MEC and NFV concepts with Small Cell virtualization in 5G networks, and enhancing them for supporting multi-tenancy. Its purpose is to provide cloud services within the network infrastructure, and to facilitate, by promoting and assisting it, the exploitation of network resource information. To this end, all the normally hardware modules of the Light data centre and the Main data centre will be delivered as resources using virtualization techniques. Both networking and computing virtualization extensions will be developed using open frameworks such as Open Platform for NFV (OPNFV). The combination of the proposed Edge data centre architecture with the concepts of NFV and SDN will facilitate achieving higher levels of flexibility and scalability.

As seen in the detailed architecture in Figure 3-6, the Main data centre will be able to execute different small cell and service VNFs under the control of the Cloud-Enabled Small Cells Manager (CESCM). In particular, the Main data centre hosts the cSD-RAN controller, which performs cRRM decisions for handling efficiently the heterogeneous access network environment, composed of different access technologies such as 5G RAN, LTE, and Wi-Fi. These radio access networks can be programmable, and are under the supervision of the centralized controller. The cSD-RAN controller updates and maintains the global network state in the form of a database called "RAN Information", which includes, among other elements, an abstraction of the available radio resources in the CESC cluster. This abstraction takes the form of a 3D Resource Grid that characterizes the resources in the domains of time/space/frequency. The RAN Information will be used by the cRRM to perform the resource allocation decisions (such as scheduling).

The cSD-RAN controller can also host centralized SON (cSON) functionalities coordinating multiple small cells, hence not appropriate for running at the Light data centre, like for example the Inter-Cell Interference Coordination (ICIC) functions. Other distributed functions like as distributed Self Organizing Network (dSON) and/or distributed RRM (dRRM) functions that are of low complexity, and that do not involve the coordination of multiple small cells, will run at the Light data centre. For example, this could be the case of an admission control function that only takes decisions based on the current load existing at a given cell.

3.2.4 Multi-access Edge Computing & Energy Efficient Radio Access Networks

Radio resource management control is one of the key tasks within the cSD-RAN controller and is executed by a set of RRM algorithms, typically consisting of a scheduler, admission/congestion control, and other mechanisms pertaining to the Medium Access Control (MAC) layer. In the case of LTE-like systems, the scheduler allocates physical resource blocks based on the traffic type (or QoS requirements) and on the underlying channel quality. In order to be able to provide effective radio resource management for the virtualized RANs (i.e. the edge cloud), with potential to virtualize various components of typical wireless stack, there are a lot of specific design challenges, which will be presented below.

The goal of such centralized resource control is on one hand assuring of the target 5G Key Performance Indicators (KPI), while considering the need of Mobile Network Operator (MNO) to reduce costs (OPERational EXpenditure OPEX, and CAPital EXpenditure CAPEX), and addressing the energy optimization requirements. Virtualization of all or portions of baseband processing of wireless stack can be centralized, enabling pooling and minimizing over-provisioning of processing resources by right-sizing. To realize ubiquitous and universal network services in various use-cases, the different solutions leverage SDN technology (two tier virtualization) to break the technology gaps and regional strictness, in both vertically tiered and horizontally celled heterogeneous networks. In this section additional (more in depth) requirements for design of centralized scheduler are presented, mainly based on [3-21] and [3-22]:

1. The use of general purpose platforms with real-time enhancements to support virtualization;
2. Cloud RAN virtualization and resource pooling;
3. Identification of effective function split between network functions (virtual or physical) at the level of 3GPP radio stack;
4. Effective interference mitigation techniques are necessary (especially as a way towards “cell less” architecture);
5. To cope with demands of 5G networks it is needed to consider the need to mitigate issues related with: network convergence, load balancing in cell networks and handover.

The above design challenges for the centralized scheduler are discussed in the following paragraphs, one by one.

Use of General Purpose Platforms (GPP) to apply virtualization: when GPP are used for RAN functions, it is challenging to ensure real-time operation so that millisecond constraints can be met for the Layer-1 function split. However it is possible to utilize GPP for other compute-heavy baseband parts such as the MAC scheduler, within highly constrained time bounds (lower than 100 us). Scheduler and MAC must execute certain set of actions in every Transmission Time Interval (TTI) – but dedicated patches may be needed, like for instance “run to completion”, which removes the randomness of the kernel interrupts to wake up tasks every TTI. Moreover, on Intel x86 processors, the Open Event Machine [3-23] (an open source framework for a user-space run-to-completion model) can be used and accelerated using the DPDK environment [3-24]. Virtual

CPUs of a virtual machine (VM) can be bound to physical ones. Applying such modifications has effect of creating an essentially dedicated environment for RAN functions.

Cloud RAN virtualization and resource pooling: as cloud-RAN enables dynamic right-sizing of the processing resources based on workload, it is essential to define the mapping between of virtualized entities (VM, containers), cells, and Central Processing Unit (CPU) cores. Especially worth considering is the potential to divide RAN modules into “per-user” and “per-cell” operations. Scheduling is “per-cell” operation because a cell’s scheduler has to consider scheduling metrics of all users in the cell (channel state and scheduling metrics of all users, and resources available) in order to make optimal resource allocations. It should be taken into account that there is a trade-off between virtualization gain and implementation complexity, for example, whether to allocate resources on a per-user or per-cell basis.

When traffic load rises and falls in cloud/centralized-RAN, Virtual Passive Optical Network (VPON) can be reformed accordingly. VPON formation enables not only resource sharing, but also Base Station (BS) coordination. The whole radio access area can be partitioned into many service areas by formation of VPONs. VPON formation enables not only resource sharing, but also BS coordination. RUs located in an area can be grouped into a VPON and controlled by the same DU.

Identification of effective split at the level of 3GPP radio stack: various options of functional splits of the radio stack largely determine the latency and bandwidth required on the fronthaul link – various splits have been analysed by NGMN [3-25], as well as in the 5G New Radio (NR) specification [3-26]. For instance, to ensure peak throughput and maximize effects of centralization, the split low at the Physical layer (PHY), breaking the Hybrid Automatic Repeat request HARQ loop, a delay of 0.5 ms is required on the fronthaul link roundtrip-time. In contrast splitting at the non-real-time Layer2/3 (such as MAC-MAC) provides larger scale centralization and lower latency/bandwidth requirements on fronthaul, while pooling gains are lowered, as baseband processing (PHY) has to be deployed near to the Remote Radio Head (RRH). The placement of particular PHY, MAC functions should be jointly optimized by the VNF Manager or Orchestrator, according to the ETSI Management and Orchestration (MANO) framework. Today, consensus is yet to be achieved on how the fronthaul traffic will be transported between RUs and DUs, and how virtualization of network resources will occur from a radio network segment to the centralized baseband processing units.

Interference mitigation techniques: with dense cell sites, the chances of interference due to poor planning will increase (more sites means more effort to optimize site parameters). Hence, various techniques for combating interference need to be considered, such as the enhanced ICIC (eICIC), Coordinated Scheduling and Dynamic Point Selection (CSDPS) for downlink, and Joint Reception (JR) for uplink, operating at low time-scale, and Coordinated Multi-Point (CoMP), operating at fast time-scale. The latency of the information exchange between cells and the bandwidth available for coordination are key aspects which impact the multi-cell coordination performance [3-27] [3-28]. The processing resources of distributed BSs in RAN (DRAN) are not designed for CoMP, and signalling undergoes long delays (4-15ms) over backhaul links connecting BSs to the core network: Centralized RAN (CRAN) solutions are thus necessary with their virtualization mechanism in place. When a user is mobile, CoMP provides seamless communication by re-forming dynamic clusters of RUs that can jointly transmit signal to the user. For resource allocation in SD-RAN, Edge cell design principles need a cross-layer optimization framework, which assigns resources in an end-to-end manner, that is, by allocating (baseband unit) processing resources, fronthaul transmission resources and radio resources for each user.

Mitigate issues related with 5G traffic demands and mobility: when the size of a cell gets smaller in 5G networks, the traffic load balance issue emerges in contrast to macro cells that can smooth the random fluctuation in the space domain. With cell size reduced to tens of meters in 5G cellular networks, quickly moving terminals lead to frequent handovers and additional latency

is inevitably added. When the handover occurs between different types of heterogeneous wireless networks, the large amount of overhead will decrease the data exchanging efficiency. The ultimate goal of designing an effective scheduler is to enable effectively “zero” interference. Such trend is recognized in literature as the “cell-less architectures” [3-29]. Proposed scheme supports the adaptive adjustment of the number of Baseline Schedulers for Access Points (BSs/APs) by the requirements of the mobile terminal and the wireless channel status in different environments.

Baseline Scheduler: a LTE eNB scheduler consisting of the two configurable scheduling algorithms, simple round robin (designed for benchmarking), and advanced channel-aware proprietary scheduling, will be used as a baseline. It is recommended to interact with LTE eNB protocol stack through a Small Cell Forum compliant interface - Functional Application Platform Interface (FAPI) extended to support Carrier Aggregation. The use of standardized FAPI interface enables out-of-the-box cooperation capabilities with popular open-source protocol stacks (such as OpenAirInterface). The criteria considered by LTE eNB Scheduler during the process of making a scheduling decision will provide a solid background to develop a centralized scheduler that considers the requirements and capabilities described above. LTE PHY Lab is oriented on 5G experimentations hence its functionalities extends the 3GPP specification by introducing potential 5G features, such as Universal Filtered Multicarrier modulation (UFMC) modulator and demodulator. Due to modular software architecture, LTE PHY Lab can be used as a framework for wide 5G exploration in our 5G projects, where other new waveforms or algorithms possible to be included in future 3GPP releases can be validated and verified. LTE PHY Lab was successfully verified and used in many projects and experiments. Currently, LTE PHY Lab is one of the key components in showcase 3 of eWINE project [3-30] where the mutual influence of Generalized Frequency Division Multiplexing (GFDM) and Orthogonal Frequency division Multiplexing (OFDM) is examined.

3.2.5 Multi-link Cooperation

5G networks are being designed to support numerous and diverse services, implying a multitude of connectivity patterns and communication paths, as illustrated in Figure 3-7 for Vehicle-to-Everything (V2X) communications. In V2X, infrastructure-based links are in fact envisioned via the Uu interface, supporting Vehicle-to-Network (V2N) communications, as well as direct links via the PC5 interface (also referred to as “Sidelink”) supporting Vehicle-to-Vehicle (V2V) and Vehicle-to-Pedestrian (V2P) message exchanges. These two links have different characteristics and consequently are associated to different features; for instance, the sidelink is expected to provide better resource efficiency, latency reduction and out-of-coverage support, whereas Uu is expected to offer higher reliability, with higher data rate. Moreover, Uu supports connectivity with remote functions and servers, as well as to local functions and servers by using local breakout, a concept wherein the data plane is routed locally through the edge cloud, without having to cross the whole core network.

The complex V2X environment might benefit from the utilization of multiple communication modes. In addition, in certain environments multiple Radio Access Technologies (RATs) are available and can be leveraged, with the additional introduction of further challenges, considering that each RAT provides different performance in terms, for instance, of reliability, capacity, and latency.

It is furthermore worth noting how the selection of the suitable communication technology should not only be driven by the QoS requirements of the related traffic; instead, the specific use case being served at the moment shall be taken into careful consideration, as well as additional information such as geographical areas of relevance of the use case together with expected vehicle’s trajectory. Such information become relevant because certain V2X use cases require reliable support for their completion once they have started. This is the case for instance of the lane merge, which is a service that regulates the speed and the trajectories of the vehicles entering a motorway and of those already driving on the motorway, in order to achieve smooth insertions.

In this case, the interaction between the network and the application should be enhanced to support such information exchange, allowing the network to choose which link and radio access technology configuration should be focused on supporting the completion of the action, as well as taking into account information on vehicle's trajectory for network tasks (scheduling, etc.).

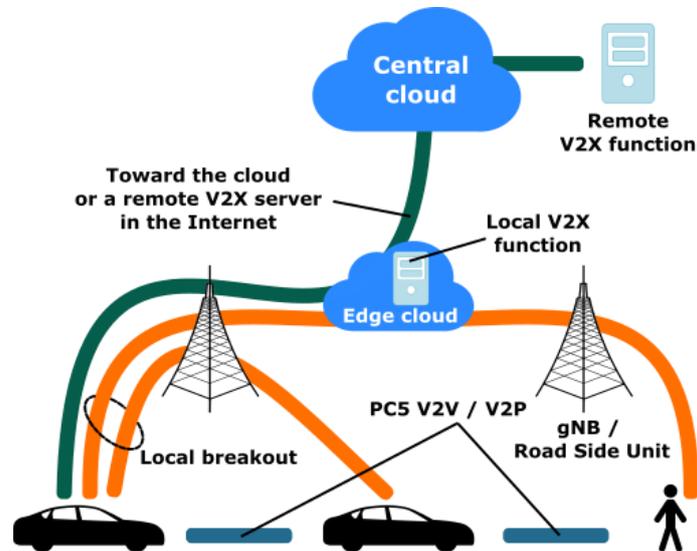


Figure 3-7: V2X communication paths

Exploiting multi-connectivity cooperation approaches, where Uu and sidelink communication modes are jointly used for several purposes, is the solution identified to address the aforementioned issues. In particular, different possible configurations have been considered, notably to improve reliability or data rate. In order to improve reliability when using sidelink as a primary link, a possible solution is to establish a secondary Uu link for redundancy in case of sidelink failure. Alternatively, both links can be used simultaneously, adding redundancy at the expense of intense resource utilization. A further example considers the dynamic selection of Uu or SL mode at the Base Station (BS), taking into consideration the QoS requirements of the V2X service and thus mapping the service to the mode (or modes) more appropriate. On the other hand, data rate can be boosted by splitting the traffic on both the Uu and sidelink, effectively increasing the amount of channel resources to which a UE has access.

The above examples can be also extended to the case of multi-RAT scenarios, thus considering the availability of different communication technologies: the aim of multi-link/RAT selection can be designed to allow the completion of a certain action or manoeuvre associated to the use case, rather than maximizing the performance on a single traffic type, or per-packet basis.

The benefits of approaches for multi-connectivity cooperation are in terms of improved reliability and data rate performance, while also providing higher resilience to link failure. Benefits might also be related to improved service availability, by jointly using several communication modes or technologies instead of relying only on one single mode or technology, which might not be able to support certain use cases on its own.

Furthermore, Road Side Units (RSUs) can be leveraged for multi-connectivity cooperation: the concept of RSU-enabled smart-zone is introduced, providing smart local radio access coverage specific to individual roads, under coordination and control of a macro radio access coverage. The smart zone, which represents an intermediate layer abstracting mobility and channel resource management, allows for efficient exploitation of Uu and SL multi-connectivity, involving both RSUs deployed along roads, and gNBs of macro coverage layer.

3.2.6 Local end-to-end Path

In many V2X use cases (like for instance cooperative manoeuvres, and sensor information sharing) the data traffic that is exchanged among vehicles (V2V) has localized significance. This means that communicating vehicles that participate in the same use case are located in the same geographical region and there is no need to access a remote server, while multiple transmission modes (unicast, broadcast, and multicast) might be required. For localized V2X communications, either the cellular (Uu) interface or the sidelink (PC5) interface could be used considering the radio conditions and the environment where the V2V use case takes place. Specifically, the NR-Uu interface could provide guaranteed QoS (i.e., high reliability, low latency) especially in cases such as no line-of-sight among communicating vehicles, poor PC5 radio conditions or high PC5 interference. Nevertheless, existing cellular solutions, based on the Uu interface, may need some updates for supporting in a more efficient way the challenging performance requirements that localized V2X services have, which include the need for fast and guaranteed transmission of localized data.

The formation of local end-to-end (E2E) radio data paths over the Uu interface is proposed to enable the fast and guaranteed transmission of localized data traffic among the involved devices, satisfying their QoS requirements and the features of the V2X services. The “end-to-end” term denotes that the (user plane) radio data paths are established among the involved communicating end devices (i.e., vehicles), while the “local” term denotes that the paths are established via the BSs. The focus of local E2E paths is that the nodes of the core network do not participate in the user plane transmissions, since the data traffic is localized and handled directly among involved BSs. Local E2E paths via the BS can support different communication modes (unicast, multicast, broadcast) without the need to interact with other entities such as the Multimedia Broadcast Multicast Service (MBMS).

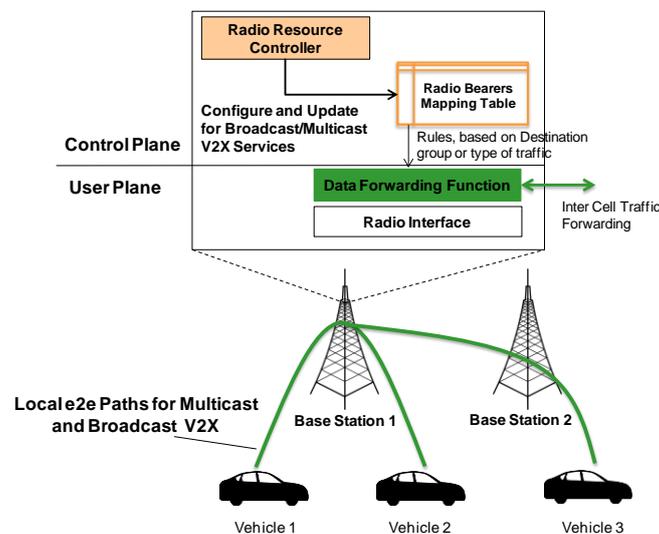


Figure 3-8: Concept of fast V2V paths via the cellular interface

Localized communication through the Uu interface requires the introduction of a data routing/forward function at the BS (gNB) that transmits the data packets among vehicles in a fast and guaranteed way. This routing table in the BS maps and connects the uplink (UL) and downlink (DL) radio bearers of different UEs for the formation of the local radio paths and consequently the faster forwarding of localized V2X traffic. According to the type of the traffic, the routing table at the BS undertakes to forward the data packet to one or more UEs in the same or neighboring cells. Figure 3-8 provides an overview of the involved entities and interfaces.

A UE requests the establishment (or update) of the local cellular V2V paths using Radio Resource Control (RRC), Non Access Stratum (NAS) protocols, for localized V2X traffic and to

transmit/receive data packets over a local E2E path. The type of the service and the identifiers of other involved UEs in the corresponding V2V service are information that the initiating UE should provide and is used for the establishment of the paths as well as for the configuration of the routing tables. RRC and NAS protocols need extension to support establishment, update and release of local cellular V2V paths between the UEs over the gNB(s) as well as to update and configure the routing table needed for the forwarding of localized data traffic. Based on these RRC or NAS messages, core access and mobility management function (AMF) and session management function can control the establishment, modification, and release of this new type of link (local cellular V2V paths) as well as to update and configure the routing tables that are introduced at the BSs in order to form V2V paths for localized V2X traffic over the Uu interface.

3.2.7 Multicast and Broadcast in RAN

The introduction of multicast and broadcast capabilities has been designed considering two aspects: the introduction of a NR mixed mode with multicast capabilities, and its extension to a terrestrial broadcast mode with similar features to the LTE eMBMS, but based on the new 5G-NR air interface.

The NR mixed mode enables a dynamic and seamless switching between Point-to-Point (PTP) and Point-to-Multipoint (PTM) transmissions: this mode is envisaged for different verticals, i.e. media and entertainment, automotive, internet of things (IoT) and public warning. The design principles of the NR mixed mode air interface design are the maximum compatibility with the NR air interface for PTP, the inclusion of essential features for PTM, such scheduling and channel acquisition for entire groups of users using a common Group Radio Network Temporal Identifier (G-RNTI), new DCI format, and a multiple cell coordination that is enabled by forcing the same cell scrambling sequence to the neighbouring cells. Negative numerologies and the concept of mini-slots are included to support single-frequency network (SFN) areas and larger deployments [3-31].

The terrestrial broadcast mode enables the reception of TV and radio services to users without uplink capabilities in mobile and/or stand-alone broadcast networks. One of its design principles is the transmission over large coverage areas in High-Power High-Tower (HPHT) networks with single-cell, multi-frequency network (MFN) and SFN configurations. To make this possible, the NR mixed mode has been extended to indicate the allocation and delivery parameters of terrestrial broadcast services transmitted in a NR frame [3-32]. A single-cell and MFN approaches have a minimal impact on the existing NR mixed mode design. The large area SFN with very large Inter-Site Distances requires new cyclic-prefix values and reference signals with a very narrow carrier spacing (even more than those provided in the NR mixed mode) in the frame structure which may also involve the definition of a new physical channel.

The 5G multicast services should be available in dynamic areas where the number of users during popular events can be high and the user distribution within the multicast area will change over the time. In the context of multicast and the NR mixed mode, the RAN is aware of UE's interest to receive data from IP multicast group. Dynamic RAN Multicast Areas (RMA) with synchronization point in NG-RAN can support multitude of deployments from a single cell DU to multiple cells under several DUs, still controlled by a single CU [3-32]. In Figure 3-9, the Intra-CU cases 1 to 3 are managed with single gNB forming the RMA for IP multicast transmission, while the case 4 reflects the inter-gNB deployment and for example terrestrial broadcast deployment: 1) Single-Cell PTM, Intra-DU, Intra-gNB; 2) Multi-Cell PTM, Intra-DU, Intra-gNB; 3) Multi-Cell PTM, Inter-DU, Intra-gNB; and 4) Multi-Cell PTM, Inter-gNB.

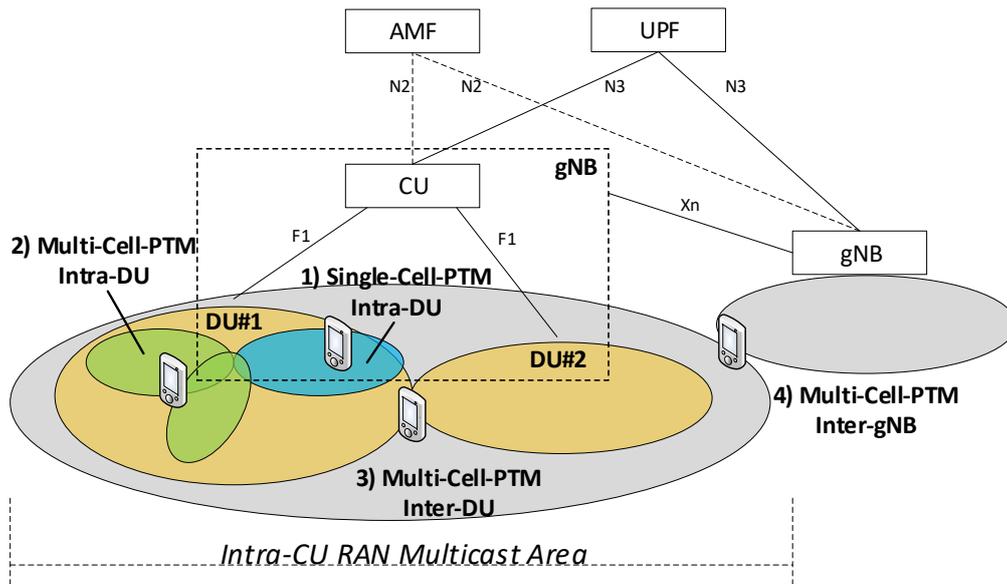


Figure 3-9: Deployments of the RAN Multicast Area, in function of single gNB

The proposed RAN architecture does not include a dedicated network entity, which functionality would include the configuration of multi-cell transmission. Instead, the approach uses run-time configuration of the transmission parameters. The content delivery over multiple cells by gNB-DUs needs a point of coordination and synchronization that ensures the same content is transmitted by gNB-DUs in time and frequency. The coordination point is hierarchically above the gNB-DUs in the RAN architecture, for instance in gNB-CU. The multicast functionality within gNB-CU may be called gNB-CU-MC and covers the user plane functions to deliver the multicast data to one or more DUs joining the multicast transmission within the RMA, see Figure 3-10.

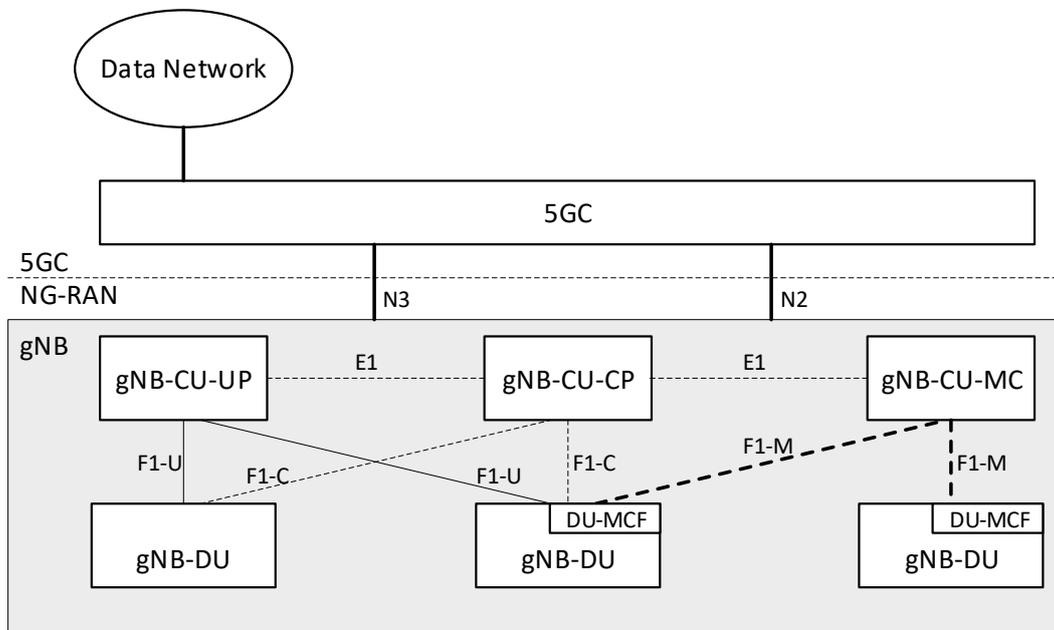


Figure 3-10: NG-RAN architecture with RAN-based synchronization of multicast / broadcast traffic

The protocols and RRM functionalities are designed to support flexibility and efficiency of new radio that is required for existing and future services including media, group communications,

and V2X communications. In this respect, multiplexing of unicast, multicast and broadcast is expected to be the factor approach in achieving the needed efficiency.

The RAT is expected to support dynamic adjustment of broadcast / multicast areas based on user distribution or service requirements and to dynamically select unicast (PTP), multicast and broadcast (PTM) transmissions. The proposed Xcast L2 protocol architecture with a switching function between Radio Link Control (RLC) and Packet-Data Convergence protocol (PDCP) layer is designed to enable these functionalities, see Figure 3-11. In cloud-based deployment, the switching function is implemented in gNB-DU. Highly flexible, efficient and reliable RRM includes [3-32] [3-33]:

- Support for flexible delivery of multicast or broadcast data via a radio-access-level seamless transition between PTP and PTM transmission modes. In a certain geographical area, if there is a limited number of UEs consuming a service in broadcast or multicast mode, better spectral efficiency can be achieved by mapping PTM radio bearers to PTP radio bearers. Moreover, if a UE that uses PTM transmission is experiencing poor radio channel conditions, transition of UE's transmission mode to PTP transmission by mapping PTM radio bearers to PTP radio bearer may improve spectral efficiency by exploiting PTP benefits such as link adaptation and HARQ (taking the latency constraint of the service into account).
- Support for selective Forward Error Correction (FEC) upon transition from PTM to PTP transmission modes. In the case of using Application Layer FEC (AL-FEC), selective FEC procedure is used to make the radio access network to intelligently select only source packets for the PTP radio bearer and both source and repair packets for the PTM radio bearer.
- Support for feedbacks for PTM transmission modes. To alleviate the heavy packet losses, which risk the technical requirements, in poor channel conditions of PTM transmission, prospect of feedback and error correction schemes have been investigated. To this end, link adaptation for PTM and a Layer 2 Error Correction (EC) in the radio access network are proposed.
- Support for use of QoS-aware feedback to optimize HARQ feedback overheads in PTM bearers. In case of very high number of UEs, ACK/NACK feedbacks can be source of extremely high signalling overhead that considerably deteriorate the network efficiency in general. To alleviate the signalling overhead to some extent, the HARQ feedbacks are optimized based on QoS requirement of the service.
- Support for efficient multiplexing of unicast and broadcast / multicast across, at least, time and frequency domains.
- Support for mechanism of link adaptation in co-ordination with higher layer error correction schemes such as layer 2 EC in order to achieve efficient and reliable broadcast / multicast wireless link that fulfils minimum expected Quality of Experience (QoE) requirements of a service at minimum cost in the spectral efficiency of the network.

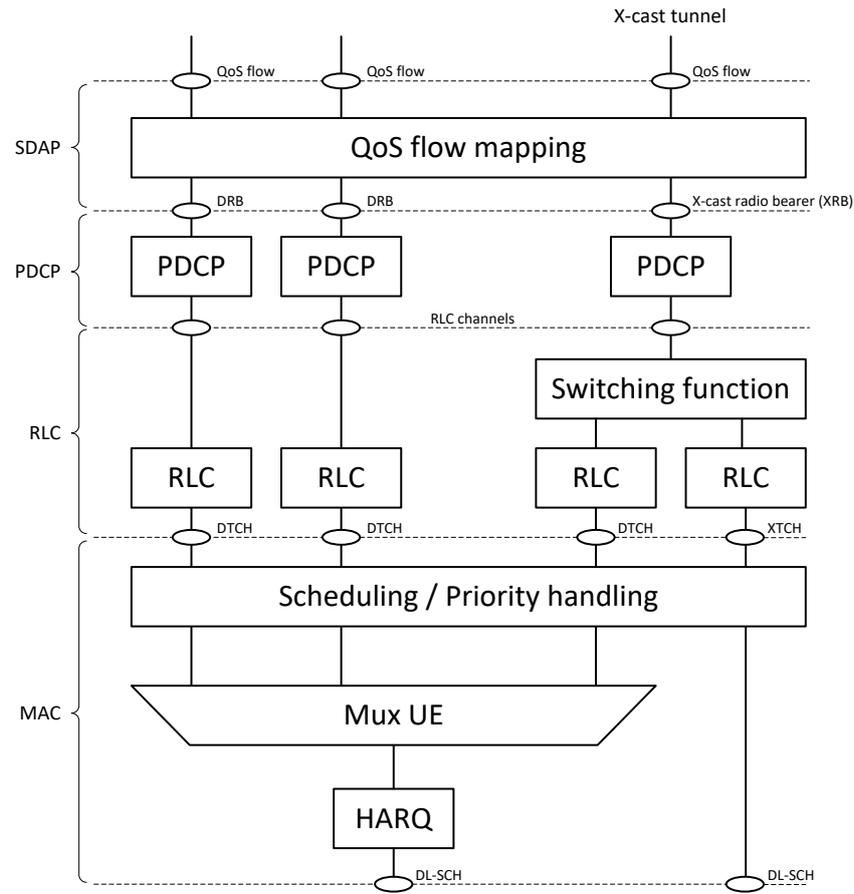


Figure 3-11: L2 architecture and bearer selection in Cloud-RAN

3.2.8 Non-terrestrial networks (NTN)

The integration of non-terrestrial networks (NTN), i.e., satellite and high altitude platform station (HAPS) based, in the 3GPP 5G eco-system aims at complementing the coverage and availability of cellular networks. NTN can highly contribute to meeting the requirements of these communities for service continuity, as well as for reliability and global coverage. Targeted end-users are the typical public end users as well as the new emerging verticals, such as transport & logistics, first responders, agriculture, utilities (energy, water distribution), mining, broadcasters, transportation (maritime, aeronautics, trains, buses, trucks or even cars) and logistics not to name others. Hence, mass market devices such as smart phones, as well as small connected IoT devices and specific directional devices are targeted.

The main cellular and satellite industry stakeholders have studied since mid-2017, the technical issues and the enabling features for 5G systems to support non-terrestrial networks. This pre-standardization work is documented in [3-41] (Study Item (SI) FS_NR_nonterr_nw on NR), [3-42] (SI FS_NR_NTN_solutions), [3-43] (SI FS_5GSAT), [3-44] (Work Item (WI) 5GSAT), [3-45] (WI FS_5GSAT_ARCH), and [3-46] (WI FS_5G_SAT_MO). Out of various architecture options that have been considered in the work, Figure 3-13 and Figure 3-14, respectively, present high-level transparent NTN-based RAN architecture (i.e., NTN communication repeats the NR-Uu interface) and regenerative NTN-based RAN architecture (e.g., gNB is implemented on an NTN platform and the NG interface signals are transferred/regenerated over the satellite radio interface-SRI) [3-42].

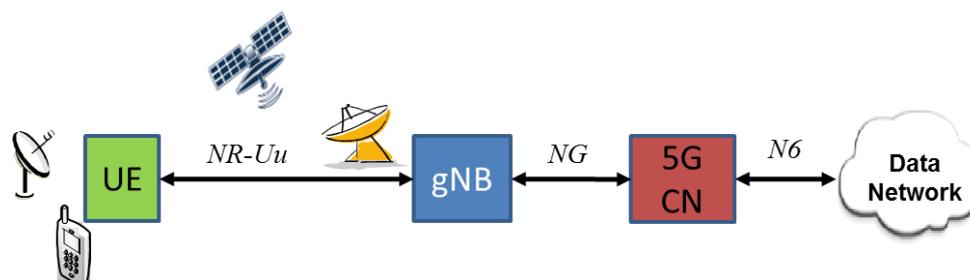


Figure 3-12: Transparent non-terrestrial platform

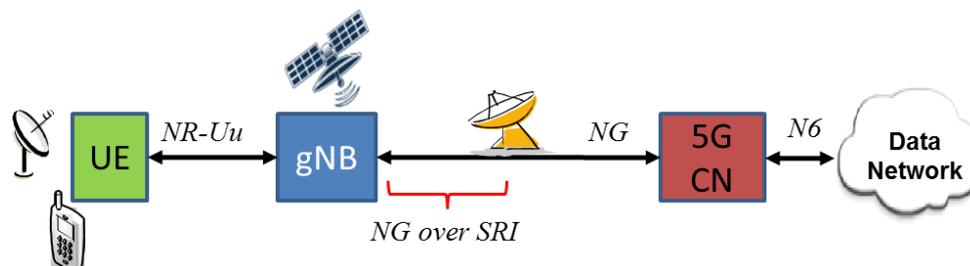


Figure 3-13: Regenerative on-board non-terrestrial platform

The next step shall consist in defining the necessary 5G features that will support non-terrestrial network solutions in the 3GPP technical specifications. NTN will be natively integrated, benefiting from technology commonalities with 3GPP defined cellular systems. As part of 3GPP Rel. 16, the 3GPP-RAN3 WG study item, “Study on solutions for NR to support NTN” and referenced FS_NR NTN_solutions, the feasibility of providing direct access from non-terrestrial platform (satellite/HAPS) to mass market smart phones is being evaluated. For mass market devices, 3GPP class 3 performances have been assumed: 0 dBi gain antenna (linear polarization), maximum 200 mW transmit power (23 dBm) and maximum 9 dB of Noise figure. The outcomes are reflected in [3-42], which is expected to be finalized by end of March 2020. Both, Geostationary Earth Orbit (GEO) and Low Earth Orbit (LEO) based NTN access are considered in the 3GPP study item as reference deployment scenarios for NTN. As examples, both S-band (~2 GHz for DL and UL) and Ka-band (DL: 20 GHz; UL: 20 GHz) are considered. While for GEO based NTN access only fixed beams w.r.t ground are considered, for LEO based access moving beams are possible as well. It should be noted that Unmanned Aircraft Systems (UAS) (including HAPS) based access has not been studied since they could be considered as a special case of non-terrestrial access with lower delay/Doppler value and variation rate.

The typical performances of non-terrestrial networks for different usage scenarios are presented in Table 3-2.

Table 3-2: Targeted performances of non-terrestrial networks (see [3-42] B.2-1)

Usage scenarios	Experience data rate		Max UE speed	Environment	UE categories
	UL	DL			
Pedestrian ¹	2 Mbps	60 kbps	3 km/h	Extreme coverage	Handheld

¹ The target performances are consolidated according to the study outcomes

Vehicular connectivity	50 Mbps	25 Mbps	250 km/h	Along roads in low population density areas	Vehicular mounted
Stationary	50 Mbps	25 Mbps	0 km/h	Extreme coverage	Building mounted
Airplane connectivity	360 Mbps	180 Mbps	1000 km/h	Open area	Airplane mounted
IoT connectivity	2 kbps	10 kbps	0 km/h	Extreme coverage	IoT

There are several effects that may create impacts on 5G standards to support “NTN”. They are listed in Table 3-3 and depend on the NTN reference scenario considered (Table 8.3-2 in [3-41]).

Table 3-3: Specific effects of Non-Terrestrial Networks scenarios compared to cellular networks

Effects		HAPS	LEO	MEO	GEO	HEO
Motion of the space/aerial vehicles	Moving cell pattern	Yes if beams are moving on earth	Yes if beams are moving on Earth (hence high speed) ²	Yes if beams are moving on Earth (hence high speed)	No	Yes if beams are moving on Earth (hence high speed)
		No if beams are fixed on earth	No if beams are fixed on earth	No if beams are fixed on earth		No if beams are fixed on earth
	Delay variation	No	High (Note 1)	Medium (Note 1)	No	Low (Note 1)
	Doppler	Low	High (Note 1)	Medium (Note 1)	Negligible	Low (Note 1)
Altitude	Latency	Negligible	Low	Medium	High	High
Cell size	Differential delay	Small	Typically relatively medium	Typically relatively medium	Possibly relatively high	Possibly relatively high
Propagation channel	Frequency selectiveness impairments	(Note 2)	(Note 2)	(Note 2)	No	No
	Delay spread impairments	(Note 2)	(Note 2)	(Note 2)	No	No

² Assuming fixed relation between beams and cells

Effects		HAPS	LEO	MEO	GEO	HEO
Duplex scheme	Regulatory constraints	FDD and possibly TDD	FDD and possibly TDD	Only FDD	Only FDD	Only FDD
<p>Note 1: Doppler and Delay variation can be pre compensated at beam centre. In such case residual Doppler and Delay variation is commensurate to the ones in cellular and can be accommodated by the UE</p> <p>Note 2: Some delay spread and frequency selective effect can be experienced in case of omni-directional antenna device especially at low elevation angle</p>						

For direct communications to mass market devices, the following should be considered:

- Operation of the satellite service link in FR1 frequency range (<6 GHz) to allow maximum commonality in the RF front end of the devices.
- FDD mode with CP-OFDM on the downlink and DFT-S-OFDM access scheme on the uplink.

Various scenario characteristics may impact the 5G system standard:

- Non geo-stationary creates Doppler shift and delay variation. This will require implementing pre compensation techniques and possibly enhanced timing and frequency acquisition, enhanced uplink timing advance/alignment procedures, adaptation of the UL transmission timing in control loops and adaptation of radio link monitoring. The need for these features depends on space segment radio configuration.
- The altitude of the non-terrestrial platforms. This impacts the need to make user and control plane protocols more delay tolerant. In particular, for high altitude (MEO or GEO), HARQ may not be relevant and hence should be turned off. For LEO, HARQ may need to be enhanced.
- Max cell size especially for LEO- and GEO-based access. This may impact the possible need for a new PRACH sequence/format/message and the adaptation of Random access protocol/procedure including 4 and 2 step RACH (RA-response window size, Rx/Tx time gap offset).
- Transparent or regenerative payload options, especially in case of moving platforms LEO, MEO and to some extent HAPS may require specific configuration of mobility procedures to support feeder link switch over.
- Earth fixed or mobile beams especially for HAPS- and LEO-based access scenarios. This may impact the tracking area and network identity management.
- UE with location determination capability (e.g., GNSS) or not especially for LEO and GEO based access scenarios. Having knowledge of the UE position, may help idle mobility (paging) as well as the uplink transmission (e.g., timing advance).
- Targeted usage scenarios (See Table 3-2). This will depends on the served UE type (3GPP class 3 or other).

Among these required adaptations (i.e., to be confirmed pending completion of the on-going study item), physical layer related ones should be also carefully studied, since they may impact the design of the chipset. However, the on-going study shows that the level of complexity of such modifications (e.g., PRACH format), if confirmed to be necessary, are likely to be very small or negligible depending on the 5G chipset design.

As a conclusion, integrating satellite components for direct access in the NG-RAN of a 5G system will be facilitated thanks to few modifications at MAC/RLC layers to cope with the latency and yet to be confirmed minor enhancements of the NR physical layer to mitigate other effects. The need for these features depends on space segment radio configuration. Hence, no or minor chipset

modifications are anticipated on the User Equipment's since the residual Doppler and delay variation resulting from pre-compensation are commensurate to the ones experienced in 5G cellular network. While future space infrastructure will support the OFDM-based NR radio interface, further RAN modifications are anticipated for existing space infrastructures to support NR. This approach should minimize the CAPEX and OPEX for the Mobile Network Operators, as the equipment will be based on a globally accepted technology, with a trained personnel to maintain and improve the software and hardware assets, with maximum commonalities with terrestrial ones.

3.3 RAN Architecture Implementation

3.3.1 RAN Deployment Options

To ensure integration with existing legacy 4G systems, and to allow independent deployments of 5G RAN and 5G Core, 3GPP has specified a set of architecture options. The numbering of options by 3GPP is not incremental in terms of the likely chronological order that deployments may occur – it is easier to consider the first likely implementation to be one of the “Non-StandAlone” (NSA) options which will then be followed by one of the “StandAlone” (SA) options.

Whilst 3GPP has defined a number of options, most implementations base their NSA deployment on Option 3/3a/3x. Options 3, 3a and 3x (Non-Standalone) allow NR deployments reusing EPC with the support of LTE eNB. With these options, the LTE eNB is connected to the EPC with Non-standalone NR. The NR user plane connection to the EPC goes via the LTE eNB (Option 3) or directly (Option 3A). In Option 3x, the solid line shown between LTE-eNB and gNB is used for user plane data transmission terminated at the gNB, i.e., S1-U data from EPC is split at the gNB. These options are shown in Figure 3-14.

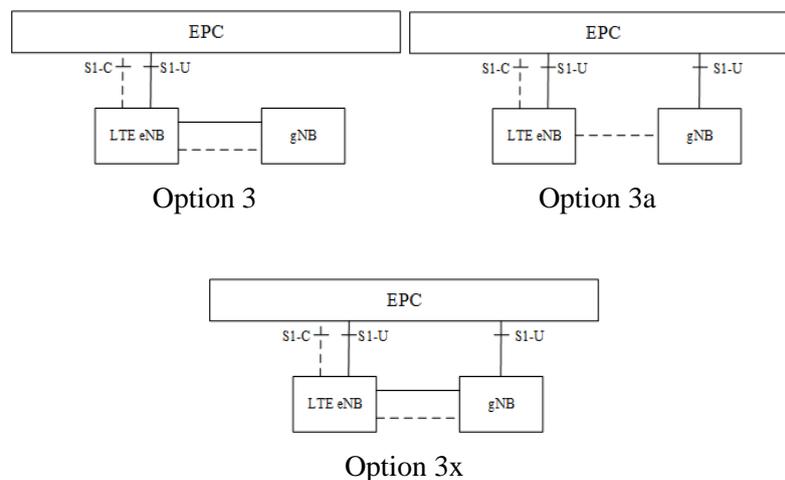


Figure 3-14: 5G System architecture, Option 3, 3a and 3x (Non standalone)

In most cases, this evolves towards Option 2 Standalone deployment. Option 2 (Standalone) defines a full 5G System, including 5G Core with gNB (Figure 3-15).

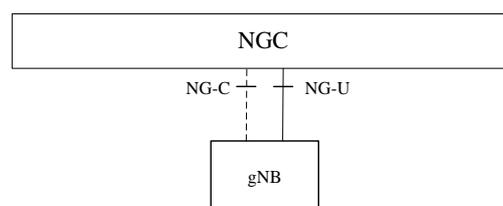


Figure 3-15: 5G System architecture, Option 2 (standalone)

The practicality of this evolutionary step may be called in to question in the future. Option 3 has a clear dependency on the presence of LTE eNodeB and EPC, and these will still be in existence even after an Option 2 5G-NR SA deployment exists. Indeed, unless Option 2 5G-NR has coverage that matches or exceeds LTE in a network, there will also be a need for interfaces between 5G and 4G systems to enable handover and fall back from 5G to 4G. As a result, whilst Option 2 may be the purist form of 5G-NR deployment, the Option 3 legacy will remain in practical terms, even if the dependency on a LTE Anchor attachment is diminished.

The deployment of new 5G radio capacity will rely on many practical factors that will cover the most prioritized requirements like the following:

- **Previous 4G radio infrastructure:** in case the radio operator already has some 4G deployed radio, the NSA options could be easier to introduce. The existence of previous radio infrastructure will be also critical for the location of the new sites for placing the new 5G radio antennas.
- **Covered area of the deployment:** we can classify the deployment in the following types:
 - Far edge: provides the smallest latency, but requires deploying the MEC services in many locations. Ideal for localized deployments like factories;
 - Aggregated edge: providing low latency, covering several radio nodes, ideal for city size deployments;
 - Regional: this deployment is ideal for services that must be provided at region level, the solution is optimal for the deployment of capacity in a regional area, covered with a few MEC servers;
 - Central: massive deployment, the new applications will be available in the whole network just by deploying a few MEC servers.
- **Frequency availability and carrier aggregation:** the different ranges of frequencies and the availability for the operators at country levels, or at the different covered areas. The bandwidth baseband capacity for the different new 5G radio frequencies and the combinations with carrier aggregation of the several 5G frequencies or 5G/4G frequencies will create very robust deployments with very high users' bandwidths in some deployment scenarios.

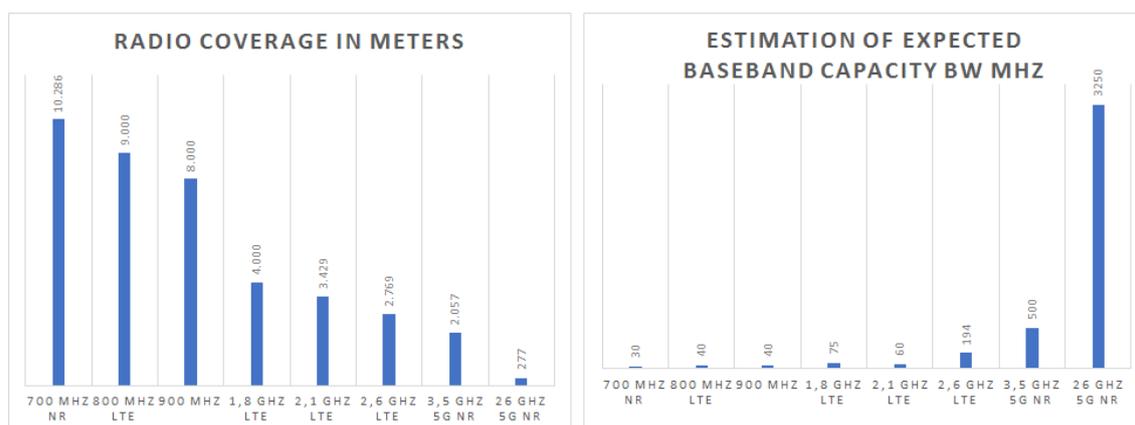


Figure 3-16: Estimation of radio coverage and baseband capacity for several 5G NR frequencies

- **MEC new applications** will support the generation of new KPIs and monitoring support by using the “Capabilities exposure” functions. In ETSI MEC, there is a specific function, namely the Network Exposure Function (NEF), to expose capability information and services of the 5G CN Network Functions to external entities, in some cases services and capabilities can be exposed over NEF, like the following:

- Monitoring: Allows an external entity to request or subscribe to UE related events of interest. The monitored events include a UE's roaming status, UE loss of connectivity, UE reachability and location related events.
- Provisioning: Allows an external entity to provision expected UE behavior to the 5G system, for instance predicted UE movement, or communication characteristics.
- Policy and Charging: Handles QoS and charging policy for UE based requests made by an external party, which also facilitates sponsored data services.

3.3.2 Visible Light Communication-based gNB

Visible Light Communication-based gNB (VLC-gNB) is a 5G small cell solution for indoor environments, as shown in Figure 3-17, consisting of two main subsystems linked together: the radio access network subsystem and the networking and services subsystem. The radio access network subsystem consists of mmWave and VLC modules which are utilizing 60 GHz unlicensed or 40 GHz licensed bands, and visible light communication to release the radio resources for the indoor environments. These technologies enable the VLC-gNB to provide Gbps data rate and sub-meter location accuracy indoors [3-34].

The networking and services subsystem consists of the Intelligent Home IP Gateway (IHIPGW). It offers intelligent management, flexible deployment, and add-on services for the VLC-gNB. The intelligence and flexibility are offered by use of SDN and VNF technologies, which enable the system to deploy UE's location server with sub-meter accuracy, which in-turn supports the deployment of add-on services such as smart TV services [3-34] location-based data access services [3-35].

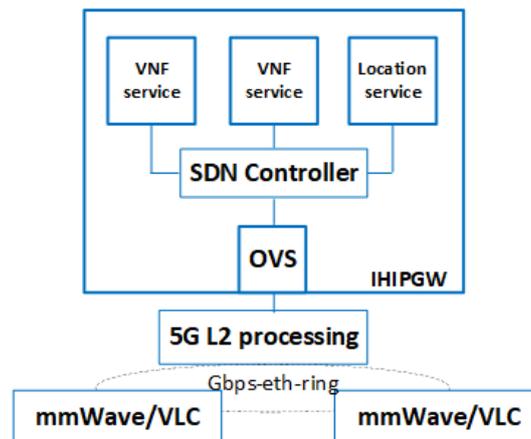


Figure 3-17: VLC-gNB small cell

The VLC-gNB provides an intelligent solution for different indoor environments such as home, museum, supermarket and tunnel stations [3-36]. It provides better QoS for UEs and offers local internet breakout, to reduce backhaul traffic, latency and improve user experience [3-35]. The next step for the VLC-gNB is to be deployed as a part of MNO RAN; however, the integration of the VLC-gNB with RAN should be considered carefully in order to provide a solution that does not downgrade the benefits gained during operation in the standalone environments. There are multiple possible deployments of the VLC-gNB indoor small cells as shown below.

Conventional topology All-Connected (AC) deployment: Each VLC-gNB small cell visible and connected back to the core network. In AC-VLC-gNB deployment, each VLC-gNB small cell is visible to the core network as shown in Figure 3-18, the UE traffic is traversed back to the core, without the involvement of the outdoor gNB. The VLC-gNB small cells use NG interface (N2/N3) to connect with 5G core while using Xn interface to connect to all other VLC-gNBs and gNBs. Adopting AC-VLC-gNB deployment makes the cost and the handover signalling relatively

high, while enables higher flexibility and lower latency in comparison to the other possible deployments.

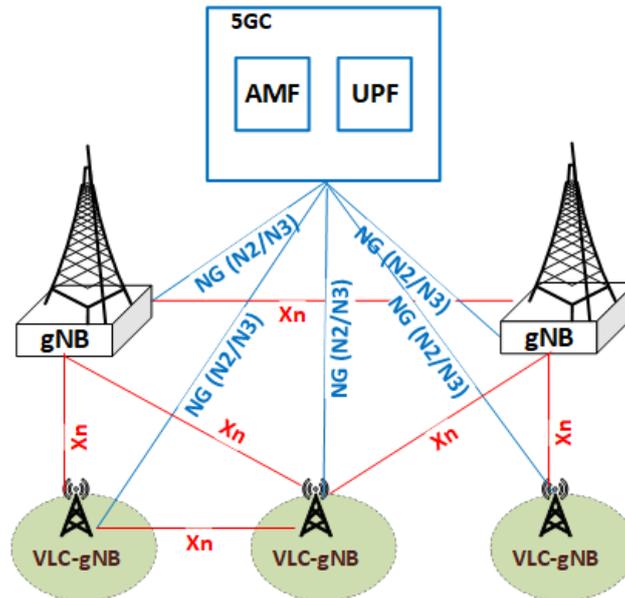


Figure 3-18: AC-VLC-gNBs topology

Dual Connectivity (DC) deployment: DC supports Stand Alone (SA) and Non-Stand Alone (NSA) deployments, the latter is considered to enable gradual transition to 5G network by enabling indoor gNB small cell to work with LTE outdoor eNB.

- **gNB and VLC-gNB DC:** UE is connected to outdoor gNB acting as a Master Node (MN) and one VLC-gNB small cell acting as a Secondary Node (SN), as shown in Figure 3-19. The MN is connected to the 5G core via NG interface and to the SNs via Xn interface.
- **eNB and VLC-gNB DC:** UE is connected to outdoor eNB acting as a Master Node (MN) and one VLC-gNB small cell acting as a Secondary Node (SN). The MN is connected to the Evolved Packet Core (EPC) via S1 interface and to the SN gNB via the X2 interface. The SN gNB might also be connected to the EPC via the S1-U interface and other SN gNBs via the X2-U interface.

Adopting DC deployment makes the cost relatively high, while enabling more flexibility, lower latency and handover signalling compared to other deployments.

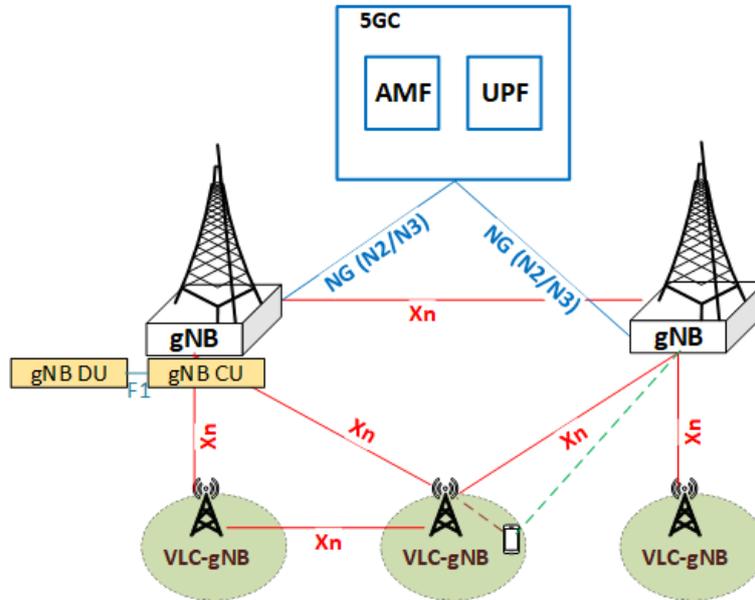


Figure 3-19: Dual Connectivity topology

VLC-gNB as Distributed Unit (DU) deployment: in DU-VLC-gNB deployment, each VLC-gNB has only RLC layer, MAC layer and Physical layer at each DU, while the Centralized Unit (CU) for a group of the VLC-gNB DUs are kept as a VNF at the gNB, named Virtual Gateway (V-GW). As shown in Figure 3-20, V-GW connects to VLC-gNB DUs using F1 interface. gNB uses NG interface to connect to 5GC and Xn interface to connect to the other gNBs. V-GW is implemented as a VNF residing within gNB to optimize the signalling and the operation of the VLC-gNB DUs, by providing one point of interaction with gNB to all connected VLC-gNB DUs. Also, it enables the VLC-gNBs to provide intelligent services since it utilizes NFV technology to offer virtualised network entities such as V-proxy/cache servers. Adopting DU-VLC-gNB deployment makes the cost, flexibility and the handover signalling relatively low, while making the latency relatively higher.

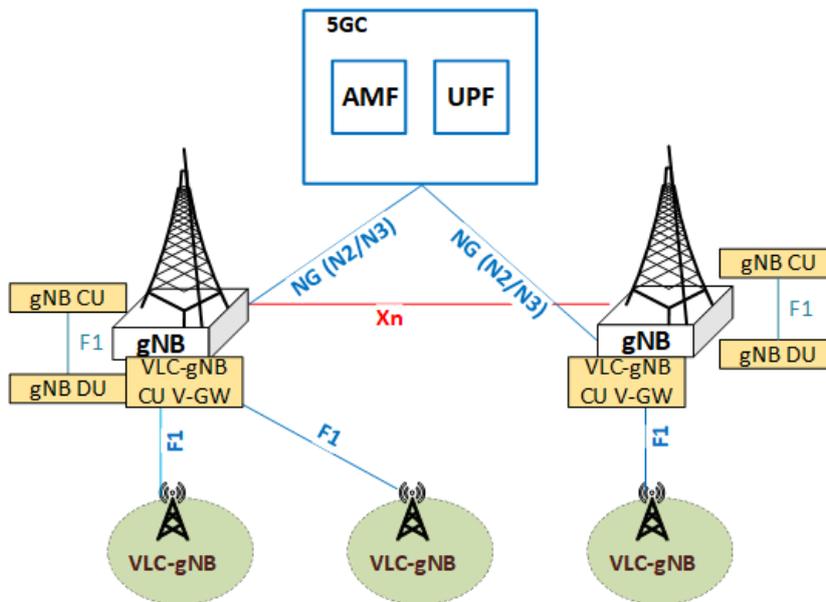


Figure 3-20: VLC-gNB Distributed Units topology

3.3.3 Lower Layer Split and Implementation

3.3.3.1 Baseband processing in active antenna distributed unit

The introduction of massive Multiple-Input Multiple-Output (MIMO) with several dozens or even hundreds of antenna elements renders current CPRI-based C-RAN architectures infeasible due to the extremely high data rates required on the CPRI Fronthaul (FH). To mitigate this, it is necessary to include parts of the Digital Signal Processing (DSP) of the baseband directly into the remotely deployed radio units. The RRH and the distributed unit become a single entity, including complete analogue and Radio Frequency (RF) processing, as well as partial baseband functionality. This is referred to as an Active Antenna Distributed unit (AADU).

Figure 3-21 shows the high-level architecture of an AADU. It can be differentiated into several Radio Sub-Units (RS-U) and an Interface Sub-Unit (ISU). The RS-Us are responsible for per antenna processing, which can include analogue and RF processing (filtering, amplifiers, data converters), calibration, and partial digital baseband processing, as well as local power distribution and an interface towards the ISU. The ISU performs all joint processing, including partial baseband processing, control & management, and an interfacing towards the CU.

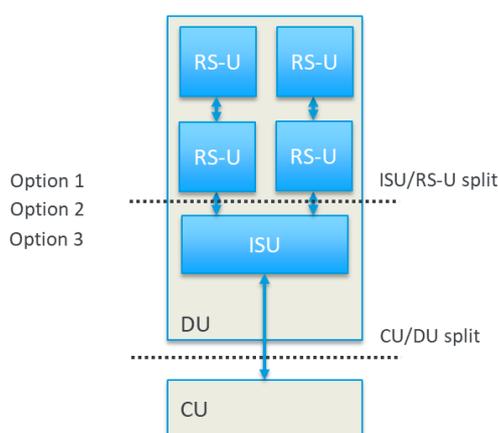


Figure 3-21: AADU overall architecture (right) and functional split options (left) [3-37].

The AADU uses a modular approach, by utilizing several identical RS-Us to compose the antenna array. This has several advantages: different antenna configurations/form factors (16x4, 8x8) can be built relatively quickly; analogue components are close to antenna element location, which avoids long RF routing distances; per-antenna processing can be performed on distributed hardware, reducing processing requirement per RS-U; finally, the data rate between the RS-U and ISU can be reduced, as each link transports data of only a subset of the antennas.

The chosen functional split has a strong impact on the design of an AADU, as it both determines the processing capabilities as well as the interface requirements. In addition to the usual functional split between CU and DU, the proposed AADU architecture incorporates an additional, inter-CU functional split between ISU and RS-U. Figure 3-21 shows three different functional split options under consideration for the AADU. In the following the corresponding characteristics are listed.

Option 1:

- Time-Division (TD) beamforming in RS-U, ISU serves only as interconnect, remaining PHY processing on CU.
- High FH and ISU/RUS data rate, but reduced compared to per-antenna transport.
- Limited processing capabilities in AADU (TD beamforming only).

Option 2:

- TD beamforming in RS-U, partial PHY processing on ISU, remaining PHY processing on CU.
- Higher computational requirements for ISU due to Fourier Transforms.
- Beamforming weights and pilots need to be transferred between CU/DU.
- Further reduced FH data rate.

Option 3:

- Partial PHY processing in RS-U, remaining PHY processing on ISU.
- Higher computational requirements for RS-U due to FFT/IFFT.
- Possibility to perform FD beamforming.
- Higher computational requirements for ISU (full PHY processing).
- GPP+FPGA required in ISU.
- Low FH data rate.

Coupled to the functional split is the interconnect architecture in the AADU, which also has an impact on the interconnect data rates. Three options can be considered, which are depicted in Figure 3-22: daisy chain, star, or column-wise interconnect. In the daisy chain architecture, data from one RS-U is passed to the next and only one RS-U is directly connected to the ISU. In this case, the final RS-U/ISU interface has to carry the data stream of all RS-U's. While this has no impact for split option 1 and 2, where the beamforming is performed at the RS-U, it could effectively quadruple the interface data rate for Option 3. In contrast, in the star architecture, each RS-U is directly connected to the ISU. This limits the data rate on each individual interface. However, it has the disadvantage of longer routing lengths, and the ISU still having to transmit and receive the full data rate. Finally, the column architecture is a compromise of the former two options, combining daisy chaining between different rows of RS-U's while using a star architecture for different columns.

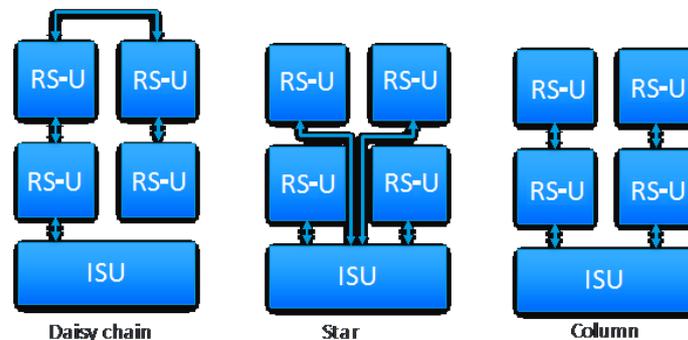


Figure 3-22: AADU interconnect options.

3.3.3.2 Optical Beamforming

The Optical Beam Forming Network (OBFN) is a photonic based technology solution [3-38], providing multiple reconfigurable RF beams at the wireless access from a single arrayed antenna [3-39]. Reconfiguration is achieved independently per generated beam, both with respect to the radiating pattern shape and the directivity of the beam. The number of the generated output beams is equal to the number of feeding (input) data streams, while their shape is determined by the number, type and configuration of the antenna elements in the arrayed antenna. The overall complexity, power consumption and footprint is significantly reduced in comparison to the single beam forming (electronic) solutions, since the signals are fed in parallel to the antenna elements (by the photonic OBFN chip). An equivalent single beam solution would require a number of arrayed antennas and driving circuits equal to the number of the generated beams.

The key idea behind the OBFN solution is based on the implementation of a Blass matrix and true time delay circuitry on an integrated optical chip (such as on the SiN TRIPLEX® integration platform). Each one of the inputs of the OBFN accepts the optical data signal to be transmitted over a certain beam by the antenna array that is attached at the output of the OBFN. The antenna elements of the attached array equal the output ports of the OBFN. Therefore, a MxN OBFN can in principle generate up to M separate beams from one set of N antenna elements. The principle of operation relies on broadcasting each one of the input optical signals to all (or designated) outputs of the matrix providing certain phase differences. Radiating multiple copies of a signal with different phase from a set of equally spaced antenna elements causes the signal copies to constructively interfere at a certain angle from the antenna array. By tuning the phases of the OBFN matrix elements the generated beams can be steered.

Two OBFN variants have been proposed for investigation in [3-40], namely a coherent and an incoherent OBFN. Incoherent processing implies that multiple wavelengths (i.e., multiple lasers) are used that are combined in the OBFN and incoherently processed in the detector. An advantage of the use of multiple lasers is the additional optical power at the photodiode and thus a better RF link performance. The key implementation drawback is that the distance between the OBFN module and the antennas must be small so that fibre dispersion will not affect the phase differences. For the coherent processing, all input data streams share the same optical source and therefore dispersion is not an issue. However, in this latter case, all the parallel output paths to the antenna elements must accurately maintain the same length so that the different copies appear synchronized at the antenna elements.

The incoherent version can apply directly at the cell site and be placed together with the antenna for both the downlink and uplink directions. An interesting solution for the coherent version relies in the combination with Multi-Core Fibre (MCF) links in the optical fronthaul distribution network. In this case the OBFN chip can be located at the Central Office (CO) part of the network, thus feeding remotely the simplified antenna array at the cell site. This is due to the fact that MCFs have an almost zero length differences between their cores and therefore can maintain (in principle) the strict phase difference requirements. However, in practice small temperature differences may cause small variations affecting the overall performance and thus require further investigation.

The initial generic designs for the coherent OBFN downlink and the incoherent OBFN uplink paths are presented in Figure 3-23. The Analogue Radio over Fibre (ARoF) scheme is adopted here as a more flexible solution for the RF spectral allocation. In the ARoF case the beam contents can share the same optical carrier and be separated at the RF spectrum domain, thus minimizing the required resources for the downlink. In the uplink direction, the received signals from all the antenna elements carrying the spectral mixture of the supported beams are driven through the OBFN. At the OBFN output, the data streams per beam are separated onto different optical carriers and can be either multiplexed and send back to the CO or sent in parallel over the MCF links. It is noted that the OBFN based multi-beam generation scheme can equally be implemented with Digital Radio over Fibre (DRoF) signals, each one occupying in this case a separate wavelength channel in the optical backhaul distribution.

In principle, optical beamforming introduces a spatial resource allocation dimension in the radio access and in addition to the RF spectrum allocation. This allows a certain RF band to be allocated to several wireless users through multiple beams from the same arrayed antenna, as long as the users are separated in space. Moreover, two or more users located at the same direction from the antenna can still receive/transmit at the peak rate of the defined standard as long as the spatially overlapping beams carry different RF bands. The combined spatial and spectral dimensions from a single multi-beam RRH optimizes the allocation of the shared resources and offers the capability to provide an increased number of connections at high data rates. The same principle of the combined spatial-spectral allocation is repeated at the optical fronthaul domain and the introduction of MCF links or more practically the consideration of fiber bundles connecting

different RRHs at the cell sites. Both options can be applicable in an architecture that relies on the parallel (spatial) transmission of signals from the CO to the antenna elements and according to the position of the beamforming integrated chip [3-40].

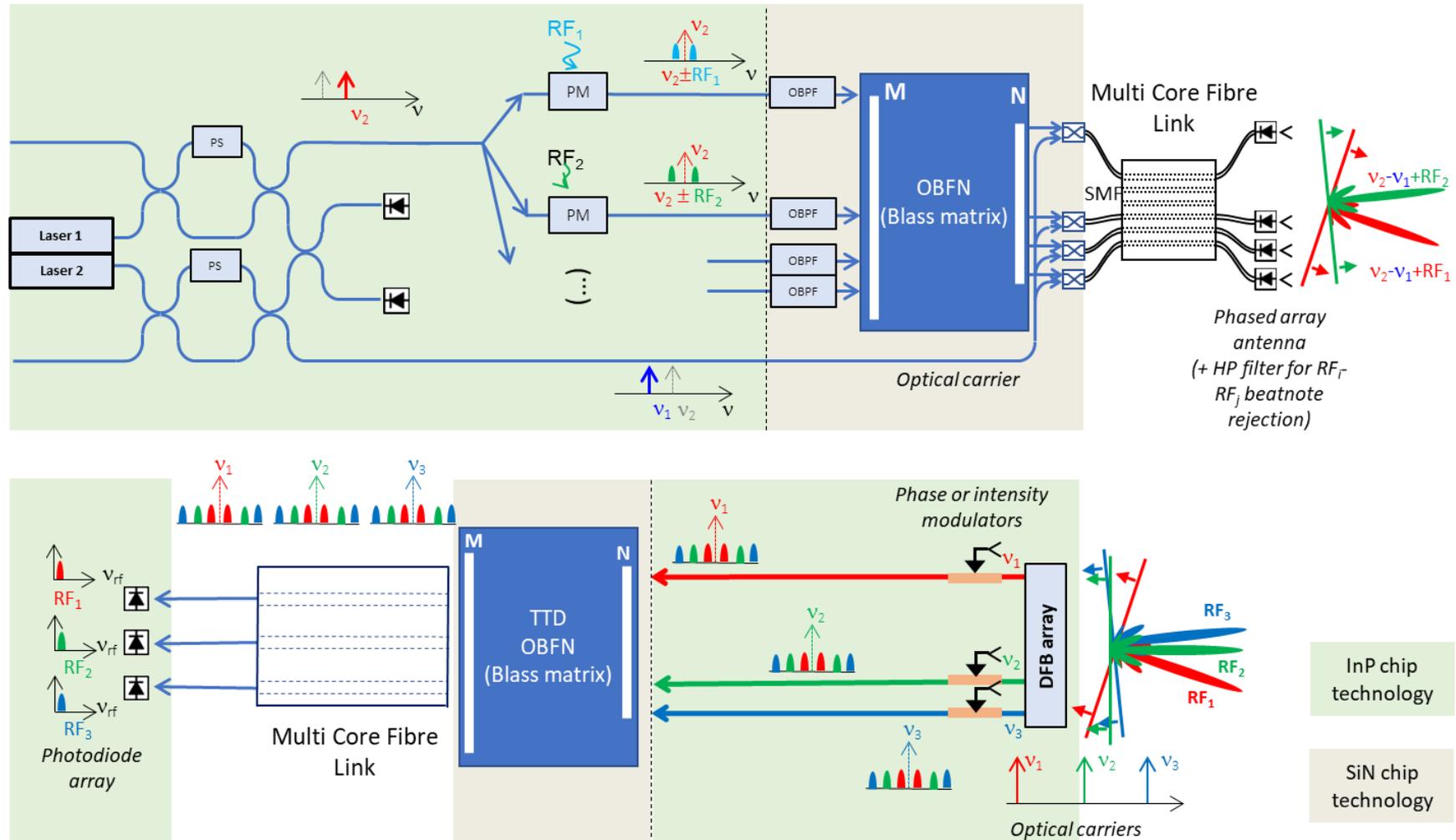


Figure 3-23: Coherent OBFN-based downlink transmitter – receiver

3.4 Edge Architecture

3.4.1 XMEC: extension of edge computing functionalities

The xMEC hosting infrastructure consists of the virtualized MEC computing, networking and storage resources, and on top of it there is the MEC NFVI. The xMEC also provides the VNFs developed and employed to enable the smart energy services. They are divided into the three groups: General VNFs, General Application VNFs and Utility Specific VNFs. The edge computing platform makes communication, computing and storage resources available for (developed) service functions of multiple-domains in an integrated way, in order to enable smart energy services. The flavour (i.e. amount of dedicated resources and scaling behaviour), (re-)location and policy for the VNFs are all managed the same way independently of the type of VNF.

The xMEC will host all complex time critical functions, as it will be in close physical proximity with the specific network element. The xMEC also provides the VNFs developed to enable the smart energy services. These VNFs can be classified in the following groups:

- General Core VNFs:
 - **vTSD** (virtual Terminals Self-Discovery): this VNF, offers device and services discovery services at a local area level (depending on the area covered by the base station hosting the xMEC stack);
 - **vSON** (virtual Self-Organizing Networks): offers device topology determination as well as optimized routing services to groups of devices that have limited network connectivity capabilities;
 - **vMCM** (virtual Machine-Cloud-Machine): allows utility resources to be stored in the cloud and accessed by multiple users overcoming any scalability issues;
 - **vMME** (virtual Mobility Management Entity): an extension to the standard LTE MME, which provides for idle mobile devices paging and tagging including GPS location (like when safeguarding the location of EVs, mobile terminals or drones);
 - **vBCP** (virtual Blockchain Processing): offers an easy-to-use and universal API gateway allowing multiple applications to benefit from the security, immutability and transparency properties of the blockchain technology;
 - **vAAA** (virtual Authentication, Authorization, Accounting): a VNF similar to the AMF, which provides services related to the administration of the field devices at the level of AAA.
- General Application VNFs
 - **vMPA** (virtual Media Processing & Analysis): performs near-real-time video stream processing and analysis so that results of the drone-transmitted video data are managed in real time, for instance);
 - **vDFC** (virtual Drone Flight Control) performs real time autonomous remote control of drones.
- Smart Energy Specific VNFs
 - **vPMU** (virtual Phasor Measurement Unit) monitors the state of the grid by measuring voltage levels and frequency values of selected grid locations;
 - **vESR** (virtual Electricity Substation & Rerouting): enables control of the local substation and electricity rerouting activities;
 - **vRES** (virtual Renewable Energy Sources): provides low-latency flexibility services to the grid operator that can be used in demand response (DR) campaigning to keep the grid balanced;

- *vDES* (virtual Distributed Energy Storage): provides energy flexibility in a certain time interval; the function deals with flexibility provisioning services to the grid operator that can use it to issue demand DR campaigning.

3.5 References

- [3-1] 3GPP TS 38.300 “Technical Specification Group Radio Access Network; NR; NR and NG-RAN Overall Description”.
- [3-2] 3GPP TS 38.401 “Technical Specification Group Radio Access Network; NG-RAN; Architecture description”.
- [3-3] 5G-VINNI Deliverable D1.1 Design of infrastructure architecture and subsystems <https://doi.org/10.5281/zenodo.2668754>
- [3-4] 3GPP TR 38.801 “Technical Specification Group Radio Access Network; Study on new radio access technology: Radio access architecture and interfaces”.
- [3-5] 3GPP TR 38.806 “Technical Specification Group Radio Access Network; Study of separation of NR Control Plane (CP) and User Plane (UP) for split option 2”.
- [3-6] 3GPP Tdoc RP-180554 “Plan for finalizing all NR architecture options”, RAN#79 meeting, March 2018.
- [3-7] 3GPP Tdoc RP-182850 “Proposals to address RAN overload”, RAN#82 meeting, December 2018.
- [3-8] 5G-MoNArch Deliverable 2.2: “Initial overall architecture and concepts for enabling innovations”, June 2018.
- [3-9] 3GPP TS 23.501 “Technical Specification Group Services and System Aspects; System Architecture for the 5G System”.
- [3-10] 5G-MoNArch Deliverable 2.3 “Final overall architecture”, April 2019.
- [3-11] CESNET, Czech Educational and Research Network, “Netopeer NETCONF server” Online at <https://github.com/CESNET/netopeer/wiki/NetopeerServer>
- [3-12] IETF, RFC 4741 “NETCONF Configuration Protocol” Online at: <https://tools.ietf.org/html/rfc4741>
- [3-13] Open vSwitch, a Linux Foundation Collaborative project, Online at <https://www.openvswitch.org/>
- [3-14] Open Daylight, a Series of Linux Foundation Projects, Online at <https://www.opendaylight.org/>
- [3-15] Open Networking Foundation, “ONF SDN Evolution”, Online at: http://www.opennetworking.org/wp-content/uploads/2013/05/TR-535_ONF_SDN_Evolution.pdf
- [3-16] Open vSwitch documentation for ovsdb, the Open vSwitch Database, Online at: <http://docs.openvswitch.org/en/latest/ref/ovsdb.7/>
- [3-17] Pateromichelakis, E. et al., "End-to-End Data Analytics Framework for 5G Architecture," in IEEE Access, 2019.
- [3-18] 5G Public Private Partnership (5G PPP) 5G Architecture White Paper Revision 2.0 – (White Paper version 2.0 December 2017).
- [3-19] Broadband Forum TR 069 “CPE WAN Management Protocol”, March 2018.
- [3-20] Broadband Forum TR-196 “Femto Access Point Service Data Model”, November 2011.
- [3-21] Agrawal, R., Bedekar, A., Kolding, T. et al. “Cloud RAN challenges and solutions” Springer Annals of Telecommunications, Volume 72, page 387-400, August 2017.
- [3-22] Xiang, W., Zheng, K. and Xuemin, Shen (editors), “5G Mobile communications”, Springer 2017.
- [3-23] Open Event Machine Development Team, “Open Event Machine: An event driven processing runtime for multicore,” Online at <http://sourceforge.net/projects>
- [3-24] DPDK development team, “Data Plane Development Kit,” Online at <http://www.dpdk.org>

- [3-25] Small Cell Forum, "Small cell virtualization functional splits and use cases" 2016.
- [3-26] Solozabal R. et al. "Design of Virtual Infrastructure Manager with Novel VNF Placement Features for Edge Clouds in 5G". In: Boracchi G., Iliadis L., Jayne C., Likas A. (eds) "Engineering Applications of Neural Networks". EANN 2017. Communications in Computer and Information Science, Volume 744.
- [3-27] Gulati S et al "Performance Analysis of Distributed Multi-cell Coordinated Scheduler," in Proc. IEEE VTC-Fall 2015.
- [3-28] Pengoria D et al. "Performance of Co-Operative Uplink Reception with Non-Ideal Backhaul," in Proc. IEEE VTC Spring 2015.
- [3-29] T. Han, X. Ge, L. Wang, K. S. Kwak, Y. Han and X. Liu, "5G Converged Cell-Less Communications in Smart Cities," in IEEE Communications Magazine, vol. 55, no. 3, pp. 44-50, March 2017.
- [3-30] eWINE, elastic Wireless Networking Experimentation <https://ewine-project.eu>
- [3-31] 5G-Xcast Deliverable D3.2 v2.0: "Air Interface".
- [3-32] 5G-Xcast Deliverable D3.3 v2.0: "RAN Logical Architecture and Interfaces for 5G-Xcast".
- [3-33] 5G-Xcast Deliverable D3.4 v2.0: "RAT Protocols and Radio Resource Management in 5G-Xcast".
- [3-34] N. Jawad et al., "Smart Television Services Using NFV/SDN Network Management," in IEEE Transactions on Broadcasting, doi: 10.1109/TBC.2019.2898159.
- [3-35] Kareem Ali, Akram Alkhatar, Nawar Jawad, John Cosmas, "IoRL Indoor Location Based Data Access, Indoor Location Monitoring & Guiding and Interaction Applications" IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, June 5th – 8th 2018, Valencia, Spain .
- [3-36] J. Cosmas et al., "5G Internet of radio light services for Musée de la Carte à Jouer," 2018 Global LIFI Congress (GLC), Paris, 2018, pp. 1-6. doi: 10.23919/GLC.2018.8319095
- [3-37] J. K. Chaudhary, A. Kumar, J. Bartelt and G. Fettweis, "C-RAN Employing xRAN Functional Split: Complexity Analysis for 5G NR Remote Radio Unit", EuCNC 2019, June 18-21, Valencia, Spain.
- [3-38] C. G. H. Roeloffzen et al., "Silicon nitride microwave photonic circuits," Opt. Express, vol. 21, no. 19, pp. 22937–22961, Sep. 2013.
- [3-39] C. Tsokos et al., "Analysis of a Multibeam Optical Beamforming Network Based on Blass Matrix Architecture," J. Light. Technol., vol. 36, no. 16, pp. 3354–3372, Aug. 2018.
- [3-40] blueSPACE Deliverable D2.2 v1.0: "Physical architecture, system and network requirements"
- [3-41] 3GPP TR 38.811; Technical Specification Group Radio Access Network; Study on New Radio (NR) to support non-terrestrial networks
- [3-42] 3GPP TR 38.821; Technical Specification Group Radio Access Network; Solutions for NR to support non-terrestrial networks (NTN)
- [3-43] 3GPP TR 22.822; Technical Specification Group Services and System Aspects; Study on using Satellite Access in 5G; Stage 1
- [3-44] 3GPP TS 22.261; Technical Specification Group Services and System Aspects; Service requirements for the 5G system; Stage 1
- [3-45] 3GPP TR 23.737; Technical Specification Group Services and System Aspects; Study on architecture aspects for using satellite access in 5G
- [3-46] 3GPP TR 28.808; Technical Specification Group Services and System Aspects; Study on management and orchestration aspects with integrated satellite components in a 5G network

4 Core & Transport Architecture

4.1 Introduction

The 5G System (5GS) architecture, as defined by 3GPP, comprises a core network (CN) and one or more access networks, e.g. a radio access network (RAN). The CN can serve mobile and converged networks and it consists of network functions (NFs), NF services and the interaction between NFs to support data connectivity and other services, while enabling deployments using techniques such as network function virtualization (NFV) and software defined networking (SDN) [4-1]. The 5GS architecture represents a logical description of the architecture.

In this context, there is a clear requirement to provide infrastructure connectivity from the Access Points (APs) to the CN, also referred to as *transport network* connectivity. Transport networks are the foundation of 5GS as they provide the network fabric interconnecting NFs, CN and RAN, and the units of RAN. The 5GS is envisioned to enable new emerging services. This also translates to new requirements on transport networks, which must evolve to meet the challenges imposed by these services. The RAN architectures discussed in detail in the previous chapter, can adopt the Cloud-RAN (C-RAN) concept requiring infrastructure connectivity within the RAN, for example between centralized units (CUs) and distributed units (DUs) referred to as *fronthaul* (FH). The C-RAN concept can overcome traditional RAN limitations, but it introduces the need to support new operational network services over the transport network to meet the challenges imposed by emerging services. The transport connectivity can be also used to support different RAN split options that allow decomposing traditionally monolithic RAN processing functions stack to a set of different units. The number of units depends on the functional split option of the RAN functions, as discussed in section 3.1.1. The allocation of functions between the RU, DU and CU, i.e., the functional split, has a major impact on the transport network and can potentially relax the corresponding requirements regarding overall capacity, delay and synchronisation requirements. The optimal split option in the C-RAN depends on a set of parameters: supported services, services requirements, technology and protocols of the FH and the backhaul (BH), etc. To maximize coordination and resource sharing gains, it is proposed to support BH and FH jointly in a common infrastructure [4-6]. This way, efficiency improvement and management simplification can be achieved leading to measurable benefits in terms of cost, scalability and sustainability.

4.2 Core Network Architecture

The support of multicast, broadcast and integrated data analytics framework in the 5GS is discussed with the reference to the 3GPP system architecture defined in Release-15 [4-1]. Firstly, we provide an overview of interfaces in the 5G CN, including the interface between the CN and RAN to help understanding the requirements on the transport network in terms of used transport protocols on these interfaces. One of the fundamental concepts in the design of the 5GS is user and control plane separations. Control NFs offer their services to other control NFs via a service-based interface (SBI) that relies on HTTP transport [4-2]. The communication between the control plane and the user plane of the CN, i.e. between session management function (SMF) and user plane function (UPF), occurs over the N4 reference point, which uses packet forwarding control protocol (PFCP) and General Packet Radio System Tunnelling Protocol User Plane (GTP-U) for control plane and user plane portions of this interface respectively [4-3]. Both PFCP and GTP-U are transported on top of IP/UDP. GTP-U is also used between UPFs and between UPF and access network. The control protocol between CN and access network at reference point N2 is NG Application Protocol [4-4], which is transported using Stream Control Transmission Protocol (SCTP) [4-5]. The transport requirements of the 5G CN are summarized in Table 4-1.

Table 4-1: Interfaces and transport requirements

Interface / reference point	Transport
SBI	HTTP
N4	IP/UDP/PFCP and IP/UDP/GTP-U
User Plane (N3 and N9)	GTP-U
N2	IP/SCTP/NGAP

4.2.1 Multicast and Broadcast

The design principles of the CN for multicast and broadcast are aligned with those of the 5GS [4-1]. Although the architecture is discussed with the focus on mobile CN, the network design is applicable to a converged network as well. Additional principles related to multicast and broadcast capabilities from a CN architecture perspective are the following:

- Enabling multicast and broadcast capabilities should require a small footprint on top of the existing unicast architecture.
- Wherever possible, treat multicast and broadcast as an internal optimization tool inside the network operator's domain.
- Consider terrestrial broadcast as a service offered also to UEs without uplink capabilities that can be delivered as a self-containing service by subset of functions of multicast and broadcast architecture.
- Simplify the system setup procedure to keep the system cost marginal. The design aims to develop an efficient system in terms of architecture/protocol simplicity and resource efficiency. Despite simplified procedures, the architecture also should allow flexible session management.
- Focus on the protocols that allows efficient IP multicast.
- Enable caching capabilities inside the network.

It is considered that multicast and broadcast capabilities of the system are accessible as part of transport or as a service, which directly relates to the envisaged NR mixed mode and NR terrestrial broadcast mode, see section 3.2.7. As part of connectivity service, multicast together with unicast belongs to a connectivity service that provides exchange of multicast and unicast PDUs between a data network and a user equipment (UE). In this solution, the UPF terminating the N6 interface and the N6 interface itself are multicast enabled. Existing NFs, such as SMF, and RAN are enhanced to enable transport of multicast PDUs in resource efficient manner. Content providers can influence how multicast PDUs are transported through the system via services offered by Policy Control Function (PCF), e.g. providing QoS requirements, in the same way as in the case of unicast. However, the system does not offer any advanced services such as reliable delivery of multicast and, therefore, this operation is referred to as transparent multicast transport [4-12]. As a service, the system offers a set of multicast and broadcast services, referred to as point-to-multipoint services [4-12], accessible via an interface with well-defined APIs such as xMB interface [4-9] [4-10]. For example, xMB offers streaming, files, application and transport-mode sessions. A content provider may control the use of associated services, e.g. reliable delivery utilizing application layer forward error correction and retransmissions, audience size measurement and metrics reporting. A geographical broadcast (e.g. file delivery, streaming or terrestrial broadcast for the distribution of TV and radio services in geographical area) can be realized using the xMB interface. The terrestrial broadcast requires only a subset of core and access networks functionalities because UEs are not required to be connected and registered to

the network. The core network functions that are involved in the provision of terrestrial broadcast services are: XCF, SMF, UPF, NRF and NEF [4-16].

The network architecture, enhanced for multicast and broadcast, is shown in Figure 4-1. The architecture introduces new NFs as well as new functionalities to the existing NFs. It should be noted that another alternative was studied in [4-13].

The UE is decomposed to an application, a converged middleware, a 5G 3GPP modem and a non-3GPP modem. In case of transparent multicast transport, multicast PDUs are passed from the modems to a network interface offered by an operating system from where they are consumed either by the application directly or by the converged middleware, e.g. an HTTP client library implementing HTTP over multicast QUIC [4-3]. In case of point-to-multipoint services, the converged middleware is a peer entity to Xcast Control Function (XCF) and Xcast User Plane Function (XUF).

The XCF functionalities related to the control plane of xMB interface include authentication and authorization of XCF for a content provider and vice versa, creation, modification and termination of services and sessions. The XCF interacts with other NF over service-based interface and over an *Nx* reference point with XUF.

The XUF is an ingress point for content from a content provider. The use plane of the xMB interface offers both pull and push options for content ingestion. The XUF functionalities include, for example, reliable data delivery over unidirectional transport (e.g. FLUTE [4-15]) and application-layer forward error correction (AL-FEC). The XUF sends multicast IP packets via an N3 tunnel to UPF.

The UPF is enhanced to support multicast group membership discovery, e.g. Internet Group Management Protocol (IGMPv4) or Multicast Listener Discovery (MLD), and multicast routing (e.g. Protocol-Independent Multicast) in addition to the functionalities already specified [4-1].

The session management functionalities of SMF are used to allocate resources and configure the system for multicast or broadcast. The session management procedures are triggered either upon a notification from UPF, a request from UE or a request from XCF [4-14].

Several steps are needed to be completed by NFs and other entities of the architecture in order to enable multicast or broadcast. Here we shortly describe the procedure related to PDU session modification that enables transparent multicast transport and results in system configuration as illustrated in Figure 4-1. This and other procedures are described in detail in [4-14]. In the first step, the UE transmits a message (e.g. IGMPv4) to join a multicast group. The reception of this message triggers a user plane event at the UPF that is notified to the SMF. The SMF searches for an existing multicast context for this group or it creates a new multicast context if the event is triggered for the first UE joining the group. Then, the SMF initiates PDU session update procedure during which the RAN receives information about the multicast group and the UE joining it. The RAN stores the information for all UEs joining the multicast group. The RAN uses the information for RAN configuration of the UEs. The RAN also decides based on this information on the most efficient transmission, i.e. a set of unicast, single-cell point-to-multipoint (SC-PTM) and multi-cell point-to-multipoint (MC-PTM) transmissions, see section 3.2.7.

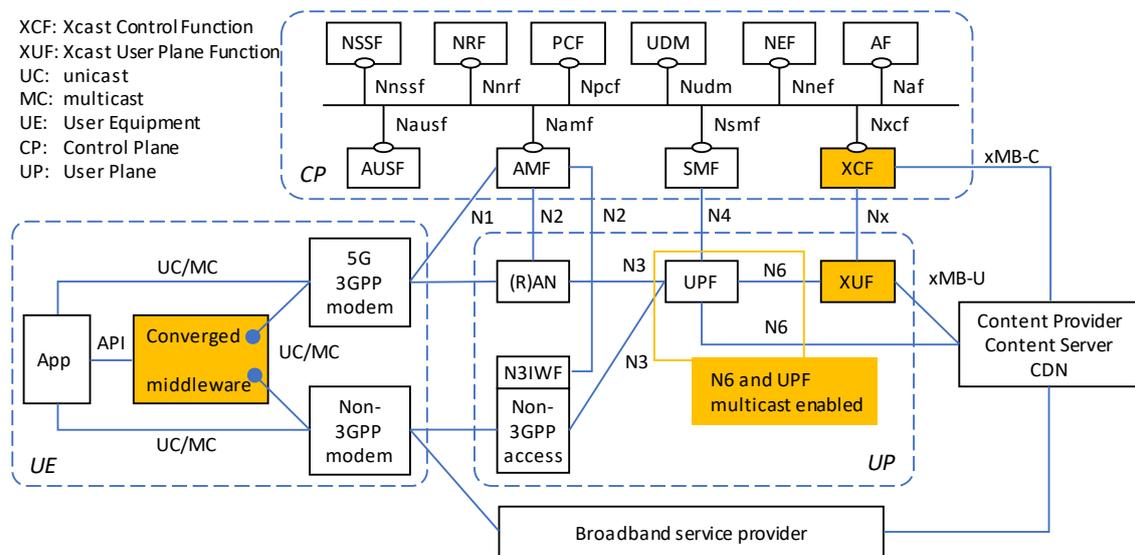


Figure 4-1: System architecture enhanced for multicast and broadcast.

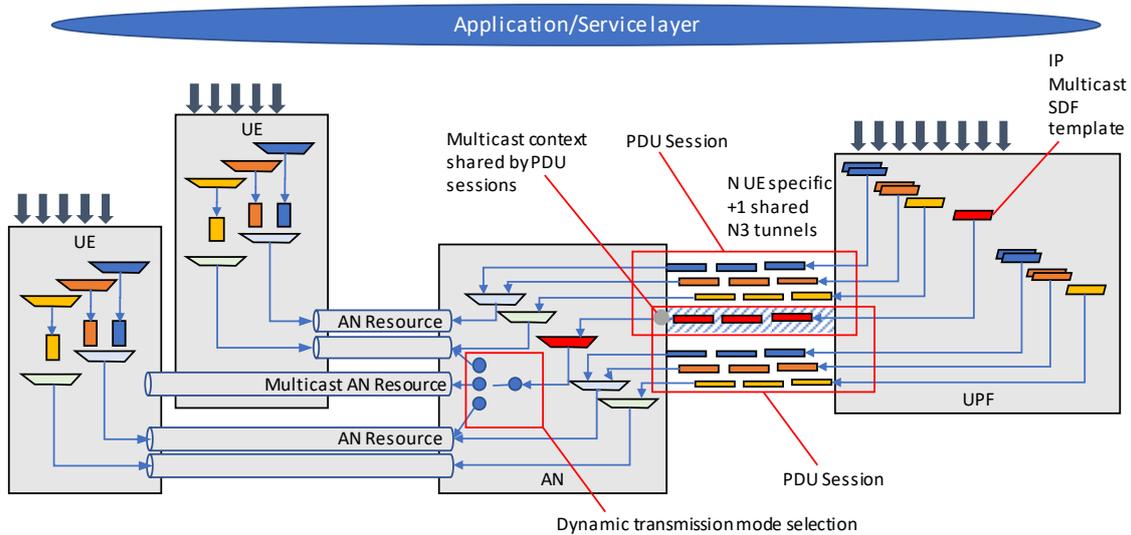


Figure 4-2: PDU Session and multicast context in 5G systems.

4.2.2 Analytics Framework

The integrated data analytics framework considers data analytics capability at various layers and introduces data analytics functions (DAFs) into core network (NW) domain (NWDAF), Big Data and Management & Orchestration (Big data/MDAF), application function level (AF-DAF), UE/RAN-DAF, and data network (DN-DAF). Each logical data analytic module is implemented as multiple instances for different use cases and purposes. For instance, the Big Data Module in the Management & Orchestration layer could be implemented as multiple instances per domains (e.g., RAN data analytics, VNF data analytics, etc.) at different levels (e.g., cross/intra domain). Such framework allows for dedicated data analytic module design at different layers, also enabling cross-layer optimisation. Different data analytics modules can be interconnected with SBIs. Below is a list of example implementation of interfaces in the integrated data analytics framework shown in Figure 4-3.

- Interface 1: NWDAF interacts with AF (via NEF) using NW layer SBI.

- Interface 2: N1/N2 interface.
- Interface 3: O&M layer configures the NF profile in the NRF, and NWDAF collect the NF capacity information from the NRF.
- Interface 4: MDAF interacts with application/tenant using northbound interfaces (NBI).
- Interface 5: MDAF interacts with RAN DAF using O&M layer SBI.
- Interface 6: NWDAF consumes the services provided by MDAF using cross layer SBI.
- Interface 7: MDAF consumes the services provided by MWDAF using cross layer SBI.
- Interface 8: MDAF collects data from NW layer via trace file/monitoring services.

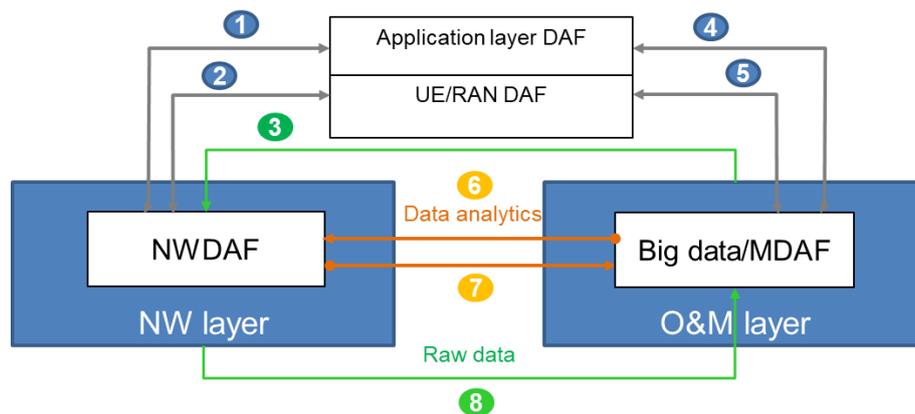


Figure 4-3: Data analytics framework in 5G-MoNArch.

4.2.2.1 Data Analytics Characterization

Firstly, we decompose the prediction/analytics functionalities in different levels, based on the predicted or expected parameter. This can involve a UE session, or the resource load/situation in a particular domain, or the application/service operation.

- **UE/Session-related parameters:** These parameters may include the prediction of the UE context/behaviour to enable the network to better provision the resources. One example can be the mobility of the user or group of users, which can be used for handover management, or the prediction of interference that the UE will suffer from/cause in a particular area. One further example is the prediction of QoS for one or more UEs in a given area.
- **Network-related parameters:** Here, these parameters can be grouped based on the domain they apply to. In **RAN**, parameters can include the ones regarding the radio resource conditions and availability (e.g., average channel quality, load, and interference) as well as the traffic (e.g., user density) and other factors in real-time or non-real time. In **transport / backhaul**, the parameters that can be estimated concern resource conditions, backhaul/fronthaul (BH/FH) type, topology, availability, dynamicity, etc. Finally, for the CN, some parameters that can be monitored are based on the processing load and availability of CN functions.
- **Service-related parameters:** This category includes the analytics which can be performed at the application domain (e.g., at terminal or at the application function) and may be used by the 5G network to improve the service operation. One example, which is specific for V2X slicing case, is the prediction of UE trajectory/route, traffic conditions, or expected Level of Automation (LoA) for a particular area.

- **Management-related parameters:** This category includes Performance Management (PM) and Fault Management (FM) analytics as introduced in 3GPP SA5. This set of parameters may take into account the current slice/subnet performance and statistics on, e.g., radio failures, and will provide analytics to the MDAF.
- **Cloud-related parameters:** This includes the cloud processing parameters, e.g., the load and availability of computational resources, which may affect the decision for virtualization of NFs to cloud platforms. In a distributed cloud-based architecture the above categories of parameters may be deployed on demand in edge or core cloud platforms. Given the tight latency and reliability requirements of some virtualized NFs (e.g., in RAN domain), performing analytics on the estimated computational resource load/conditions is of key importance for performing actions, like offloading the processing load to other cloud processing units.

Granularity of analytics

Real-time: The analytics can be performed in real-time operations (e.g., channel prediction in ms time scale). However, this is a more challenging task due to the fact that additional processing might be required and the overhead may affect the performance.

Near-real time / Non-Real time: In this case the analytics is performed in sec/min/hour time scale and may apply to certain types of prediction (e.g., load distribution in a geographical area). In Open-RAN (O-RAN), near-real time operations have been defined to capture operations like QoS management, traffic steering, mobility management, which may be semi-dynamic (e.g., 100 s of ms timescale).

On demand: This can apply to both real-time and non-real time analytics, and is the case when the vertical or the operator requires enabling this feature as a service, for a given area or time window in order to meet the requirements of a network slice.

Type of analytics

There are different types of analytics that can be useful for the network according to the Gartner's Graph on stages of data analysis [4-48]:

- Descriptive Analytics – Explaining what is happening now based on incoming data.
- Diagnostic Analytics – Examining past performance to determine what happened and why.
- Predictive Analytics – An analysis of likely scenarios of what might happen.
- Prescriptive Analytics – This type of analysis reveals what actions should be taken.

4.2.2.2 Integrated Analytics Architecture

In this section we discuss architecture enhancements and functional design considerations. Hence, the front-end is explicitly described as placeholder for employing analytics. The actual processing and data mining (e.g., what type of predictors or algorithms are used on top of these functionalities) and whether this involves multiple iterations and interaction between different entities is not shown, since this is an implementation specific aspect. Our intention is to prepare the grounds in 5G architecture for supporting analytics in multiple levels with different objectives, while these can be consumed by any authorized functionality in a slice-tailored manner. The necessity for new analytics functionality in the 5GS may become a reality, preferably using a service-based architecture, since both network operators and verticals may easily deploy analytics on demand. For example, analytics functions may be realized as (part of) a new AF, which can closely interact with, e.g., CN functions using SBI, or as CN/RAN functions which can interact with MDAF using the control-to-management interfaces.

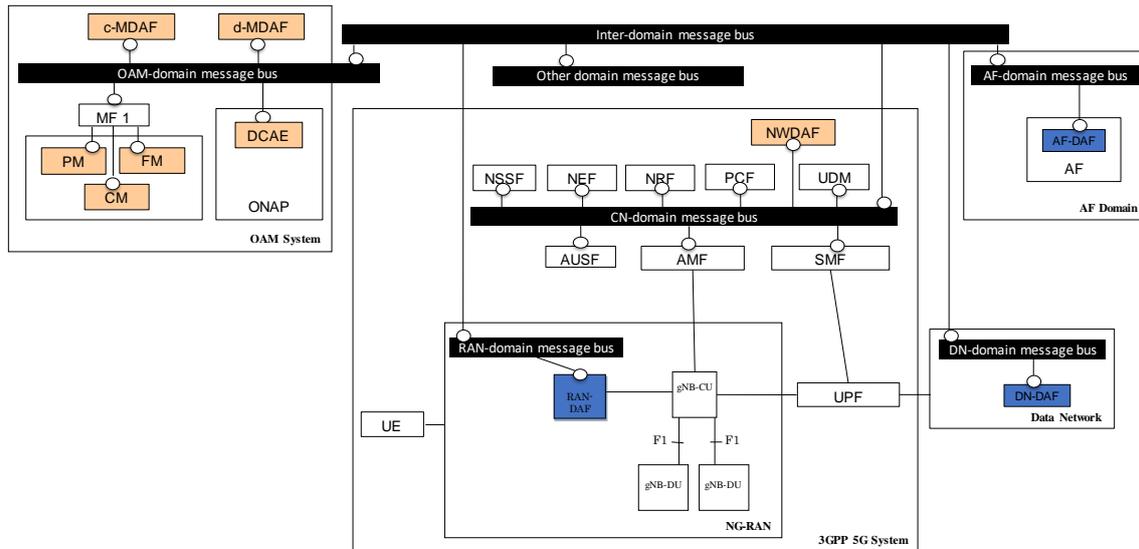


Figure 4-4 Integrated Analytics Architecture.

The functionalities that can be defined as necessary parts of the E2E analytics design framework are provided by RAN-DF, NWDAF, MDAF, AF-DAF and DN-DAF. AF-DAF and DN-DAF, as shown in Figure 4-4, and they can be deployed in two relevant domains outside the 3GPP 5GS, i.e. in AF domain and DN domain respectively. In the DN, the network operator or the vertical can place functions that provide data related to service or performance of non-3GPP networks (e.g., metropolitan wide area networks, wide area networks) to other DAFs within the 5GS or an Operations, administration and management (OAM) domain. AFs or dedicated AF-DAFs can interact with the CN-domain NWDAF, either via 3GPP Network Exposure Function (NEF) or via an inter-domain message bus, as depicted in Figure 4-4. AF-DAFs enable the operator to deploy on demand new functionality customized for AF-domain requirements, or the vertical to perform analytics that can support the E2E service operation. This can prove highly beneficial for vertical industries like IoT and V2X, where the vertical requires exposure of selected data from 3GPP network operation, a higher level of control of the network, as well as flexibility of deployment.

Real-time analytics are required for improving RAN NFs, like radio resource management. Therefore, the real-time analytics are collected from real-time measurements and are used locally for optimizing performance dynamically. Also, the business aspects may involve different stakeholders among RAN, CN, and Management. The RAN analytics may be abstracted to CN or OAM. An example deployment of such functionality is shown in [4-49], where more complex RAN deployments with CU-DU splits, better motivate for such functionality. RAN-DAF could be presented in the system architecture either as a control function in RAN (shown in Figure 4-4) or as a part of a management / SON functionality. RAN-DAF interfaces other network functions via the inter-domain message bus interface.

Intra- and inter-domain message buses provide the functionality for registration, discovery and consumption of services within a domain or across domains. Service registration and deregistration allow a service catalogue function to maintain an updated list of services available for consumption. Service discovery functionality allows to retrieve available services, refer requesting consumers to them and provide the means to access them. Service consumption functionality allows consumers to invoke services, e.g., by automatically routing requests and responses between service consumer and producers. This may include platform-like functionality, such as, load balancing, failover, security, message delivery rules, or protocol conversion / adaptation, and exposure of services to the inter-domain message bus and its service catalogue.

4.3 Transport network Infrastructure

4.3.1 Wireline technologies

One of the main challenges the transport network needs to address is to provide connectivity between DUs and CUs using commonly digitized formats. These are already standardised or under standardisation, such as the Common Public Radio Interface (CPRI) and the enhanced CPRI (eCPRI), adopting more flexible interface options in the RAN that allow to adapt functional splits between CUs and DUs to use case requirements as well as transport network capabilities. In addition to the digitized FH solution, there is significant attention to analogue FH solutions that try to take advantage of reduced complexity Radio over Fibre (RoF) solutions, offering, on the other hand, reduced architectural and connectivity flexibility.

The variable service requirements of 5G RANs demand the development of novel solutions at the transport network for the interconnection of the RUs with distributed (MEC-type) or centralised compute resources (Data Centres) for the processing of the BB functions. These solutions must provide, at the same time, high levels of flexibility, resource and energy efficiency. The adoption of high capacity and flexibility transport networks that rely on scalable and energy as well as cost efficient programmable technologies represent a future proof approach to address the challenges imposed by current and future RANs.

In this context, the joint consideration of network softwarization and programmability of advanced hardware (HW) solutions will allow a variety of tasks to be dynamically allocated between centralized and distributed elements. This will enable physical and virtual network functions to be appropriately combined and deployed on top of any programmable compute and/or network element. As an example, programmable NFs can be either placed locally at the network nodes (suitable for low-latency applications), or at a remote server (no strict latency constraints).

In the following sections present some of the most advanced transport technologies that are currently being tested in different European 5G testbeds.

4.3.1.1 Programmable Elastic frame-based Optical Transport

To address the required frequent network reconfigurations, advanced optical network solutions are adopting very flexible and dynamically changing network architectures and technologies addressing directly the nature and characteristics of services in terms of data types, traffic flows generation, and end-to-end connectivity requirements [4-17]. However, apart from flexible architectures and dynamically adopting technologies, optical networks need to support programmability features to match the very diverse and rapidly changing high bandwidth connectivity requirements of the 5G network. This programmability takes advantage of active and elastic optical technologies that can be programmed and controlled according to service level requirements. Although commercially available solutions perform optical switching supporting wavelength switching granularity, the very diverse requirements of operational and end-user services demand new approaches. These approaches would deploy more dynamic and flexible solutions to offer higher sub-wavelength level granularity, together with more elasticity in terms of optical spectrum.

One such example of an active optical transport is the Time Shared Optical Network (TSON). This frame-based optical network solution offers sub-wavelength switching granularity [4-18]. TSON provides connectivity between RUs, DUs and CUs and can enable the concept of virtual BBUs (vBBUs) and facilitating efficient sharing of compute resources.

TSON supports Elastic bandwidth allocation at a time frame level. Although natively TSON allows handling of Ethernet frames, its configuration can support a broad range of framing structures and communication protocols including CPRI, eCPRI and Open Base Station Architecture Initiative (OBSAI), either natively or through their packetised versions. To achieve

this, TSON relies on a generic and flexible resource allocation framework adopting a hierarchy of three levels of resource granularity including connections, frames, and time-slices (Figure 4-5). Connection refers to a sub-wavelength light path establishment between any two end points in the TSON domain. Each frame is divided into time-slices as the smallest units of network resource, i.e. the actual sub-lambda resources. The frame length and the number of time-slices inside a frame define the minimum granularity achievable by the TSON network. The TSON framework offers a very flexible optical platform that supports sub-wavelength switching, frame lengths, varying from 64 ns to 25.6 μs and variable bit rates, spanning from of 30 Mb/s up to several Gb/s, with 30 Mb/s step.

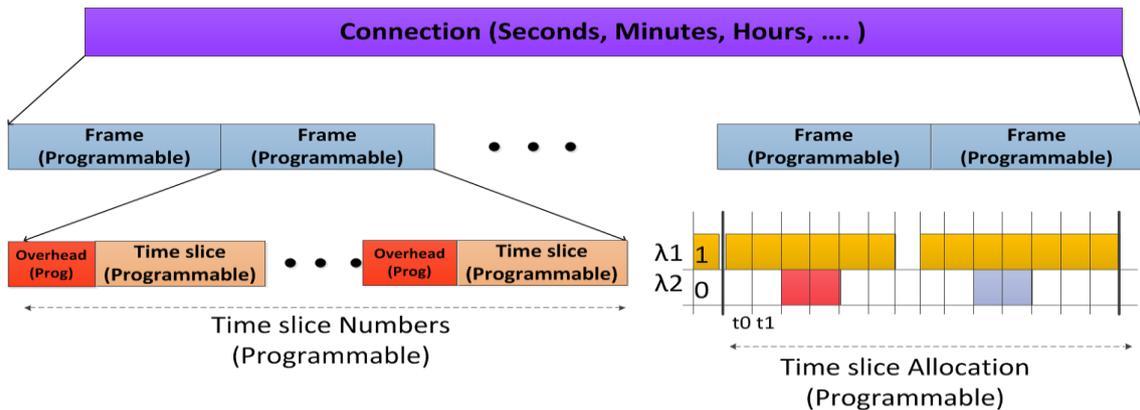


Figure 4-5: Structure of connection, frame and burst.

The TSON solution includes two different types of nodes, the **edge** and the **core nodes**, incorporating different functionality and level of complexity. TSON **edge nodes** provide the interfaces between other technology domains, including RAN, Passive Optical Networks, and MEC or data centre domains hosting compute/storage resources and vice versa. A typical example is shown in Figure 4-6, where Passive Wavelength Division Multiplexing (WDM) provides flexible FH connections between RUs at the antenna side and BBUs at the central office (CO).

The edge nodes represent a hybrid subsystem able to handle continuous (I/Q streams) and packetized flows (Ethernet traffic). The optical bandwidth allocated to the different services is not fixed but can be elastically defined based on the requirements of each service. The TSON **core nodes** do not carry out any data processing but switch the traffic optically. Therefore, the FPGA-based TSON core node controls fast-optical switches to setup the path according to the service requests.

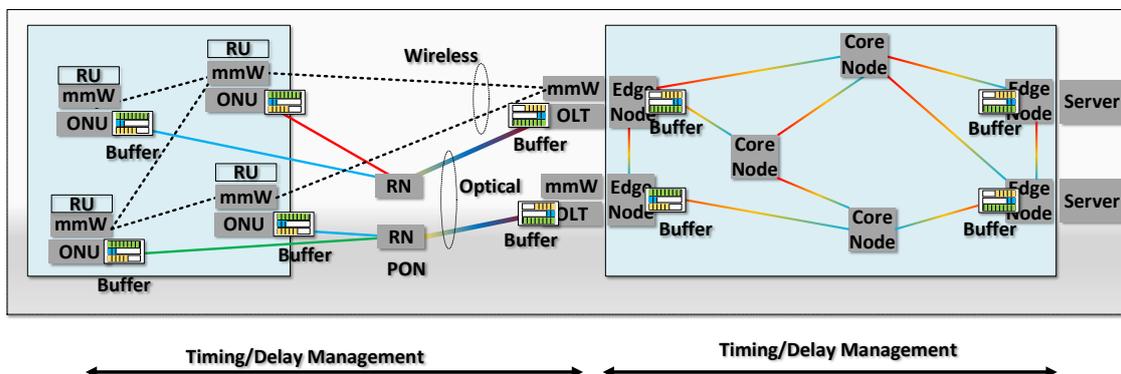


Figure 4-6: TSON integrating RAN, PON and compute domains.

TSON provides also the ability to multiplex eCPRI and CPRI traffic by appropriately assigning the suitable resources (wavelengths, timeslots), and setting different priorities for different traffic flows depending on the needed QoS. An example of this process is shown in Figure 4-7 where

two Ethernet-based eCPRI flows are aggregated by TSON Edge 1 into one flow and are then multiplexed together with the CPRI flow, assigning a different wavelength to each of them. These two wavelengths are then fed into a Wavelength Selective Switch 1 (WSS 1). Then, WSS 1 multiplexes eCPRI and CPRI packets over a single fibre and sends them to WSS 2. The WSS 2 receives the upstream flows and demultiplexes them into eCPRI and CPRI packet flows based on their wavelength. The TSON Edge 2 node receives the packets from the Ethernet and CPRI ports and passes them individually to their clients. In the downstream scenario, the reverse operation is performed.

TSON is fully SDN enabled, and the parameters of TSON nodes are programmable by a suitable controller. In addition, TSON supports programmable traffic flow control (i.e. VLAN, Dest MAC, Src MAC).

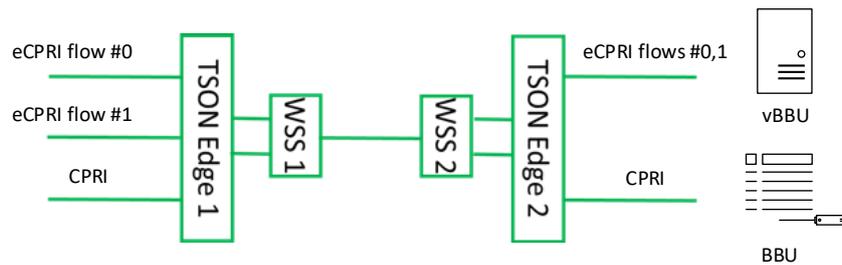


Figure 4-7: Multiplexing of eCPRI and CPRI over TSON.

4.3.1.2 Ethernet Transport

The reduction of overall costs and the improvement in operation efficiency represent two of the key objectives for both SDN and carrier Ethernet. On the transport network “all IP, the all Ethernet” technology is aiming to provide the underpinning over which the MNOs will build their future programmable networks, based on SDN and being slice-ready.

Furthermore, Ethernet bridges were originally designed for best-effort traffic with no requirement on maximum delay through a network. Due to the need of using Ethernet for audio and video transport in professional studios, there has been a drive in IEEE 802.1 Ethernet standardization for mechanisms ensuring zero congestion packet loss, as well as control on delay and Packet Delay Variation (PDV). Recently, main drivers for further evolution in standardization include industrial control and automotive applications, with mobile FH as the most recent.

For the transport network the main activities are around Ethernet over Multi-Protocol Label Switching (EoMPLS), Ethernet over SONET/SDH, Ethernet over DWDM and Ethernet over Optical Transport Network (OTN). Recently, approaches focus also on Flex-E and X-Ethernet technologies for higher capacities and to handle load dynamicity and provide performance guarantees.

Flex-E technology [4-19] is introduced as a thin layer, known as Flex-E Shim, which lies between Ethernet MAC and Physical Coding Sublayer (PCS), as depicted in Figure 4-8. The Flex-E Shim layer is responsible for the mapping of Flex-E clients (i.e. Ethernet flows) to groups of PHYs; the time multiplexing between client groups; and driving the asynchronous Ethernet flows over a synchronous schedule over multiple PHY layers. Using Flex-E, the MAC layer speed of a client can be decoupled from the actual PHY layer speed, while multiple MAC clients over multiple PHY layers can be supported even for data rates out of the conventional range offered by current Ethernet standards. Flex-E can run on top of an OTN-WDM-based PHY.

In particular, each aforementioned layer supports:

- **Data Link Layer:** a) Logical Link Control (LLC) for multiplexing network protocols over the same MAC, Media Access Control Sublayer (MAC) for addressing and channel

access control mechanisms, and Reconciliation Sublayer (RS) that processes PHY local/remote fault messages.

- **PHY Layer:** a) PCS performs auto-negotiation and coding, b) PMA sublayer performs PMA framing, octet synchronisation/detection, and scrambling/ descrambling, and c) Physical Medium Dependent Sublayer (PMD) is the transceiver that is physical medium dependent.

Each Flex-E client has its own separate MAC, RS above Flex-E shim which operate at the Flex-E client rate. The layers below the PCS are used intact as specified for Ethernet. As a first step in every Flex-E client flow, a 64b/66b encoding is performed to facilitate synchronisation procedures and allow a clock recovery and alignment of the data stream at the receiver. Then a procedure of idle insert/delete is performed. This step is necessary for all Flex-E clients in order to be rate-adapted, matching the clock of the Flex-E group.

The rate adaptation is accomplished by idle insertion/deletion process, according to IEEE 802.3. This rate is slightly less than the rate of the Flex-E client in order to allow alignment markers on the PHYs of the Flex-E group and insertion of the Flex-E overhead in the stream. Then all the 66b blocks from each Flex-E client are distributed sequentially into the Flex-E group calendar where the multiplexing is performed. An introduction on Flex-E and possible Flex-E use cases are described by Google in [4-20] and [4-21]. In [4-22] the authors present an integration approach of control and management of Flex Ethernet over OTN.

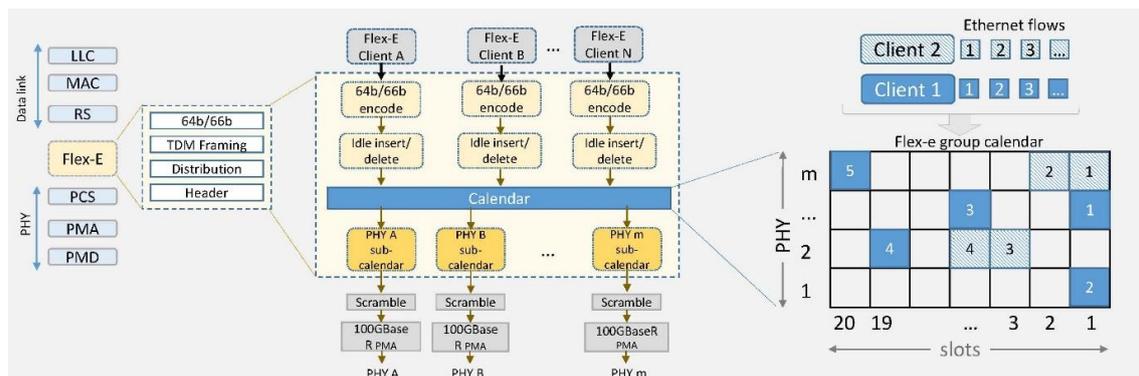


Figure 4-8: Flex-E layer between Ethernet MAC and PCS. Additional FlexE Shim distribute/aggregate sub-layer in PCS/PMD.

Furthermore, Time Sensitive Networking (TSN) Ethernet mechanisms have recently been proposed for enabling low latency in Ethernet combined with statistical multiplexing. TSN is the set of IEEE 802 Ethernet sub-standards that are defined by the IEEE TSN task group. The new standards describe several mechanisms for improved or even guaranteed real-time delivery of Ethernet traffic.

Besides the relevant IEEE activities and the activities in IETF in the Detnet WG, proprietary solutions also exist to enable a low-latency transport FUSION technology [4-33], involving a Guaranteed Service Transport (GST) class with ultra-low PDV. This enables high accuracy synchronisation by timing transparent transport of IEEE 1588 - Precision Time Protocol (PTP) - packets. In Figure 4-9, an Ethernet FUSION TSN network for aggregation, transport and de-aggregation in FH is illustrated, while Figure 4-10 shows how the Ethernet FUSION-TSN may be further combined with WDM aggregation enabling a scalable FH transport.

SDN and programmable Ethernet transport networks are currently being investigated in and within the scope of an overall programmable data plane framework, where the relevant control systems are managed by an integrated 5G orchestration and management solution.

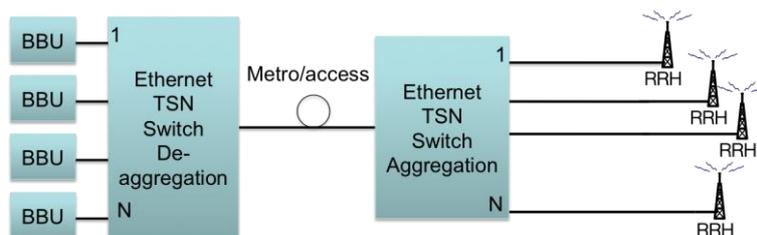


Figure 4-9: Ethernet TSN network for aggregation, transport and de-aggregation in FH.

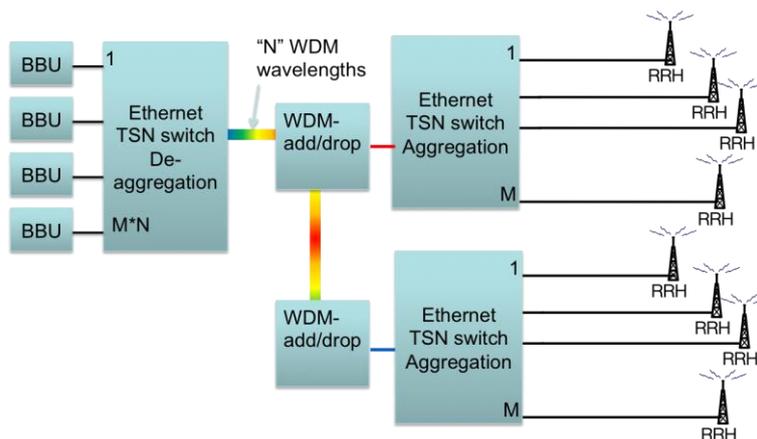


Figure 4-10: Combination of Ethernet TSN with WDM aggregation enabling a scalable FH transport.

Furthermore, a programmable Ethernet solution is useful for adapting the required functionality to the requirements of the variable functional split options. The different types of FH functional splits define different requirements with regards to latency, bitrate and traffic pattern. For example, a CPRI over Ethernet mapping will produce a Constant bitrate (CBR) stream of data at a high bitrate compared to the offered user data-rate. The eCPRI splits, on the other hand, allow statistical multiplexing and lower bitrates for the same offered user data-rate while latency requirements remain as strict as the ones for CPRI. The latter are further relaxed for higher level splits. Hence, a programmable Ethernet transport may accommodate an adaption of functionality needed for meeting the different requirements from the different functional splits.

4.3.1.3 Programmable Metro Network - Disaggregated Edge Node

The design of cost-effective, energy-efficient, agile and programmable metro networks is a relevant subject nowadays. As well, scalability has to be kept in mind when designing such a network. Some of the expected features of such network are [4-23]:

- Design of all-optical metro nodes (including full compute and storage capabilities), which interface effectively with both 5G access and multi-Tbit/s elastic CNs.
- Interconnection by novel, spectrally efficient, and adaptive optical transmission networks.
- Implementation of advanced concepts, e.g. HW disaggregation and virtualization of the disparate elements of transmission, switching, networking, compute, and storage, orchestrating dynamic solutions for multiple 5G applications.

Figure 4-11 depicts the disaggregated central office architecture. Compared to the legacy central office, this architecture adopts NFV and SDN technologies, allocates computing and storage resources closer to customers, and provides dynamic, on-demand and cost-efficient services for 5G use cases.

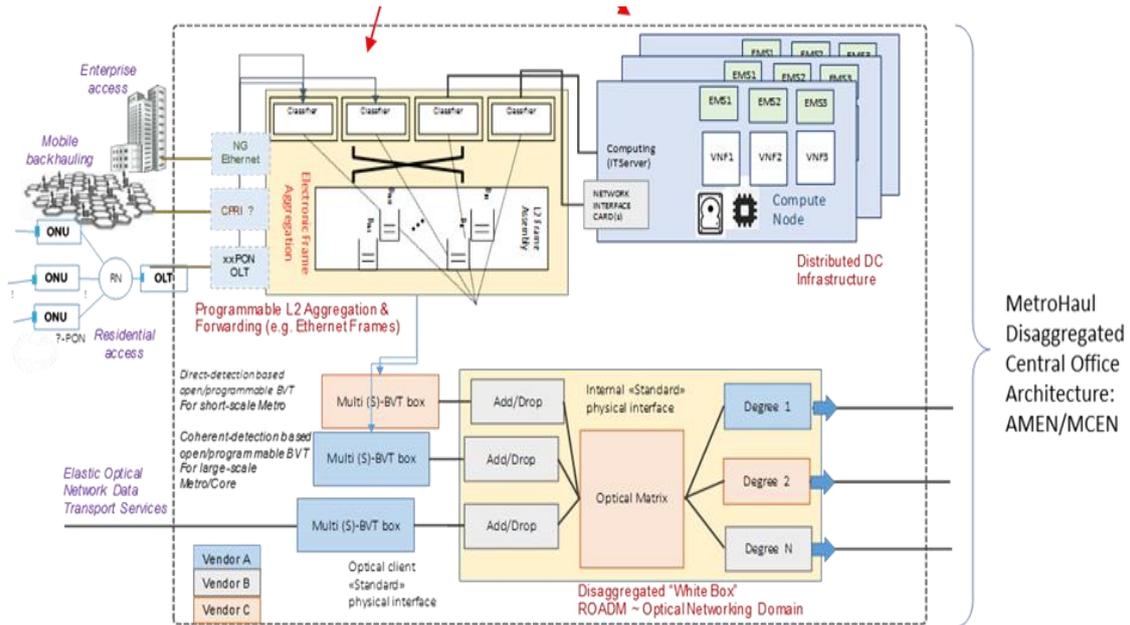


Figure 4-11: Disaggregated Metro-Haul Central Office in Network Service Platform.

The disaggregated edge node can be programmed on demand to support multiple FH/BH protocols. They are SDN enabled and can aggregate/disaggregate any access traffic combination (i.e. Ethernet, Wi-Fi, LiFi, eCPRI, etc.) to/from either TSON (metro) or coherent (core) optical networks on demand. Figure 4-12 shows the Disaggregated Edge Node architecture comprising the TSON technology, Voyager, and WSS.

TSON, referred above, is the first multi-protocol programmable interface that meets 5G Key Performance Indicators (KPIs) [4-23]. The **Voyager** is a Broadcom Tomahawk-based switch with added DWDM ports called Voyager acting as a disaggregated optical transponder [4-25]. It supports PM-QPSK, 8-QAM, and 16-QAM modulation formats. The **WSS** provides filtering and switching of the optical signal [4-28]. The control plane is composed of an SDN controller and device agents that allow to program and dynamically configure the different components of the disaggregated edge node.

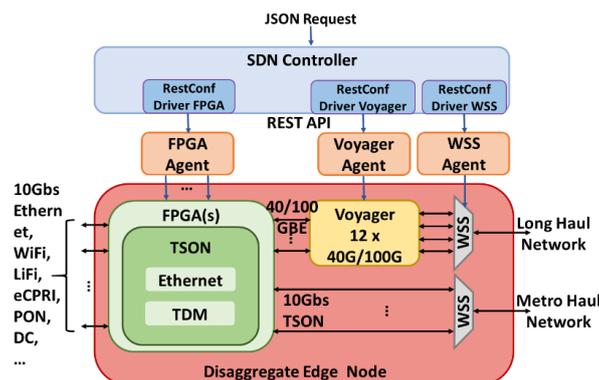


Figure 4-12 Disaggregated edge node architecture

4.3.1.4 Space Division Multiplexing

In view of the support of densely located 5G small cells for applications in crowded areas, smart offices and industrial network environments, in which a large number of closely located users or devices seek high bandwidth access and support of advance services, advanced transport topologies and architecture models are required. Space Division Multiplexing (SDM) can

significantly help increasing the amount of aggregated capacity that can be carried from the attached RUs and processed at a large centralized pool of base-band units (BBUs).

The overall concept relies on the use of an optical FH Infrastructure architecture for the connections between the pool of BBUs at the Central Office (CO) and the served RU at the cell site (CS) that considers: a) Multiple single mode fibre (M-SMF) links, that form today the majority of the fibre infrastructure investments by all major operators; or b) Multi-core fibre (MCF) links, which are the compact high capacity alternative for future capacity expansions in optical networks. Each of the above poses different characteristics to the network design by determining the type of technology solution that can be implemented, as well as the future expandability in terms of capacity and cost. The overall design principle considers small-cells that are separated either in the spectral or spatial dimension. The Optical Distribution Network (ODN) can have fixed paths to the various CS or include dynamic wavelength or spatial add/drop nodes for a segment of served CS.

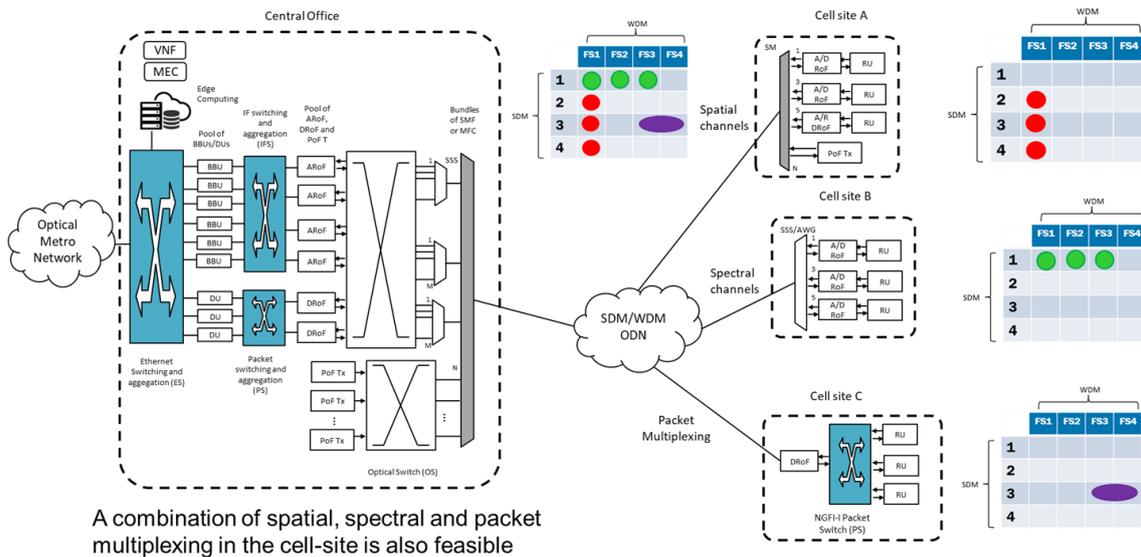


Figure 4-13 SDM-enabled infrastructure architecture

The two dimensions define an allocation strategy that may span over space and spectrum and can even be handled separately by the different CSs. The complexity though is transferred at the CO part and the management of the BBU (for pure A/RoF) or DU (for DRoF) pool of elements and the interconnection with the ODN outputs.

Technology considerations

The use of the SDM FH with legacy DRoF connectivity can provide increased capacity (or equally higher densification of cells), as well as the potential use of the space dimension for the dynamic optimization of the allocated resources according to the user demands. Further benefits can be exploited with the use of A/RoF transmission in the mmWave band and in combination with advanced optical beam forming (OBF) processing. Many antenna elements per RRH can be addressed by utilising parallel fibre links over the SDM-fibre connections. In addition, the adoption of the optical beam forming network (OBFN) element enables the simultaneous multi-beam transmission of data streams by the same antenna arrays [4-26]

To control the allocation of spatial and spectral resources over the transport infrastructure (data plane), some key enabling solutions are developed and described below.

Spatial/Spectral resource discovery and network topology information is provisioned and handled by a hierarchical Transport SDN control approach with modular capabilities. Three child SDN controllers are introduced for the FH, BH and NGFI segments, while one parent SDN controller

on top is acting as the FH/BH transport network controller. For the handling of the data plane resources on the FH segment, SDN node agents are deployed at the cell sites (CS) and the central office (CO) (aggregation node or edge) that use NETCONF protocol with the child SDN controller. At this point, the management of the physical network functions (PNFs) is provisioned as part of NFV network services for mobile communications. For this reason, a new component, called PNF Agent, is introduced in the SDM architecture as a driver to interact with specific PNFs mainly physical RRUs and BBUs. A PNF manager is also required to interact with the orchestrator. Moreover, multi-tenancy and network slicing are adopted to optimize the usage of the physical infrastructure through virtualization and resource sharing techniques, while guaranteeing high levels of flexibility in the provisioning of dedicated services with customized QoS.

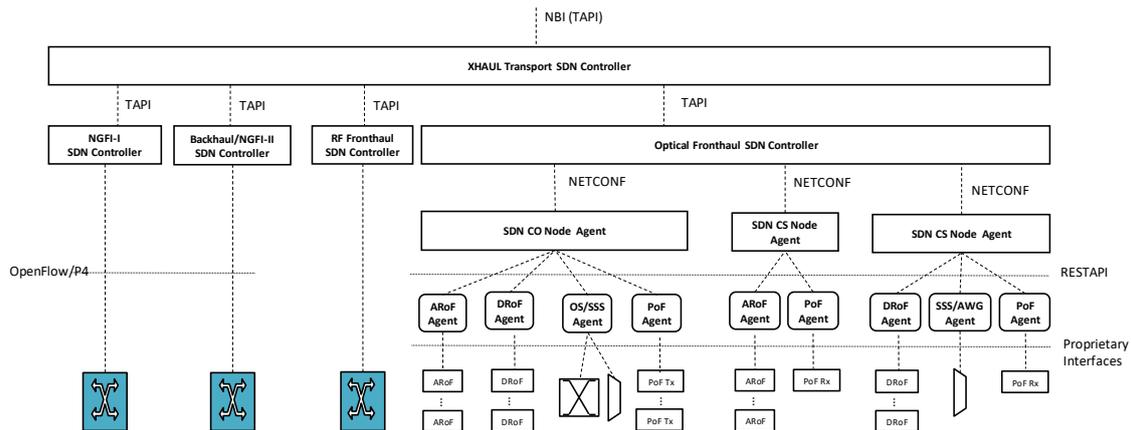


Figure 4-14 Hierarchical transport SDN architecture [4-27]

For the planning of the physical network resources, a split is applied between the RF spectrum related resources in the access part and the spatial-spectral optical resources in the FH. The NR15 parameters are implemented and can be extended with the capabilities of the SDM-assisted ARoF technologies that are applied. The optimum allocation may consider various optimization criteria like minimize overall consumption (by reducing the number of active nodes) or maximize end user performance. The initial planning (offline allocation) is important for the identification of the appropriate number of resources to handle the expected demands and especially with respect to the number of required BBUs and transceiver modules in combination with the supported total number of RRH at the cell sites and the served connections. An online planning option is also considered for the dynamic reallocation of resources among the wireless access sites as the demands vary. Programmable FPGA based BBUs and optical transceiver integrated chips, as well as dynamically adjusted beam steering solutions at the OBFN chips are implemented and controlled via the PNF agents.

4.3.2 Wireless

Network heterogeneity in 5G involves the integration of advanced wireless systems, allowing the interconnection of a large variety of end-devices. The wireless transport and access network will be based on Sub-6 technologies, mmWave technologies and massive MIMO techniques using much greater numbers of antennas at the gNBs to improve data rates, reliability as well as energy efficiency [4-37]. These will coexist with legacy (2-3G), Long Term Evolution LTE (4G) and Wi-Fi technologies to allow broader coverage and availability, higher network density and increased mobility.

From a wireless technology perspective, the transport network is currently considering a dense layer of small cells operating in the frequency range of 100 MHz – 100 GHz. Seamless integration of mmWave BH technology with Sub-6 Non-Line-of-Sight (NLoS) technology is generally

recognised as the technology providing the ideal combination of capacity and coverage by operators deploying wireless BH, particularly in complex urban deployments [4-37].

Additionally, we consider the use of satellite communications as part of the 5G network acting as a transport network that provides connectivity between areas.

4.3.2.1 Millimeter wave (mmWave)

Millimeter wave presents itself as a key technology to address the increased data rates required for serving dense urban areas. Additionally, the combination of high data rate and high-resolution ranging represents one of the key features this technology can unlock in the near future [4-38]. This feature can be useful for new services such as safety critical applications, augmented reality, assisted living, etc.

mmWave wireless BH links are currently established using multi-gigabit meshed BH technologies based on WiGig (IEEE 802.11ad) operating in the V-band at 60 GHz. Such solutions make use of electronic beam-steering to establish different topology configurations. These nodes are expected to support beam-tracking and Multiple-Input Multiple-Output (MIMO) techniques, and are enhanced with programmable network processors to allow network functions to be easily configured/modified or controlled by an SDN controller.

These technologies are currently coexisting with Sub-6 GHz technologies to reap benefits from their co-location [4-7], as Sub-6 solutions allow NLoS operation and can complement mmWave nodes in situations where mmWave nodes face NLoS conditions. Sub-6 technologies will be provided with self-backhauling capabilities. The co-location of these technologies can as well benefit the localization of additional stations [4-39]. As an example, Sub-6-GHz angle measurements can support mmWave beam pointing mechanisms using wider beam widths to a degree that the latter can allow precise localization of static and mobile nodes.

A key feature of mmWave solutions is the combination of optimized hardware accelerators with programmable parallel processing. Namely, being both MAC and PHY software-defined, allowing the performance of novel mmWave wireless algorithms to be explored and continuously tailored in the context of advanced research platforms [4-40].

Yet another specific work in mmWaves is the combination of multi-antenna (MIMO) techniques at these frequencies [4-33]. Concretely, the mmWave Line-of-Sight (LoS) MIMO architecture is especially interesting for wireless BH applications, where very high data rates need to be supported. The number of parallel data streams supported by these systems is determined by two factors, the antenna array arrangement and the wavelength-transmission range product [4-41]. This means that the spacing between antenna elements is correlated with the achievable link range (i.e. the separation between transmitter and receiver). In other words, when additional streams are needed, either array size has to be increased or, wavelength or range has to be decreased.

The maximum achievable rates for different antenna configurations are shown in Figure 4-15. These results represent theoretical upper limit on the achievable rate. For practical systems, this rate is lower and limited by RF impairments. For example, considering the IEEE 802.11ad standard, with the estimated SNR of 25.95 dB, the highest modulation and coding scheme (MCS12) the data rate of 4.62 Gb/s could be supported. That means that the aggregated data rates for a system represented with indices 1, 2, 3, 4, 5, and 6 would be 18.48, 27.72, 36.96, 41.58, 46.20, and 73.92 Gb/s, respectively.

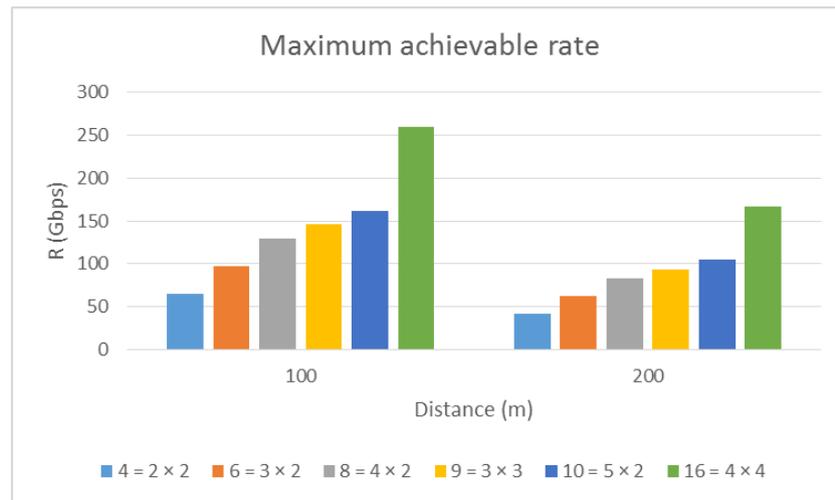


Figure 4-15: Maximum achievable rate of the 60 GHz LoS MIMO system for different antenna configurations at 100 and 200 meters distances.

4.3.2.2 Multi-tenant small cells with IAB

The development of novel technologies that support the massive deployment of outdoor small cells are required to fulfil the 5G promises on capacity. A key technical enabler is the concept of multitenant small cells with Integrated Access and Backhaul (IAB) support [4-42]. An implementation of this concept is presented in this section.

The proposed technology allows an infrastructure operator that manages a small cell deployment, to dynamically instantiate connectivity services, i.e. virtual networks, on behalf of its tenants (i.e., MNOs). The provisioned virtual networks allow the MNO's customers to connect to the small cells in a transparent manner, and carry the customer's traffic to each MNO's core network, while supporting mobility. Figure 4-16 depicts a deployment scenario where two MNOs provide connectivity services over a shared small cell infrastructure.

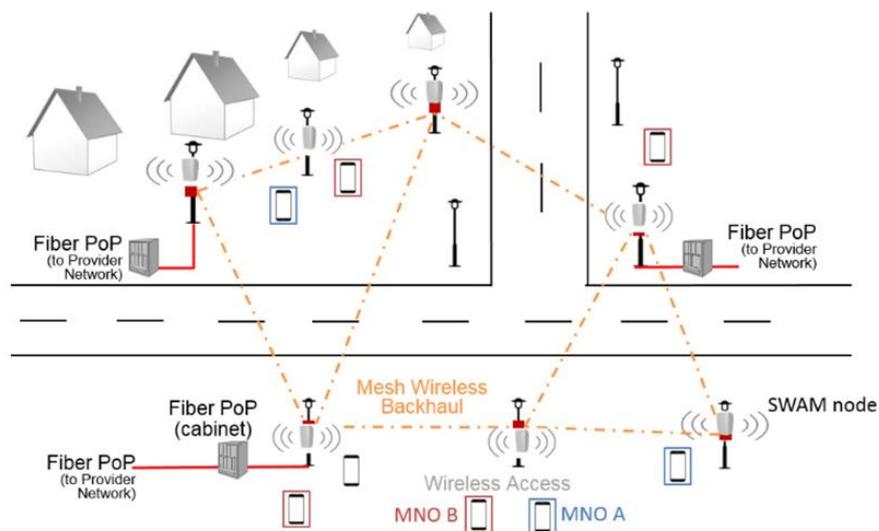


Figure 4-16: Deployment scenario for the joint access-backhaul function.

Although the proposed IAB architecture is independent of the Radio Access Technology (RAT), hereafter, we describe an implementation tailored to IEEE 802.11 technologies, which we refer

to as *SWAM*: “*SDN-based WiFi Small Cells with Joint Access-Backhaul and Multi-Tenant capabilities*”.

The services provided by *SWAM* can be divided in: i) Instantiate an access connectivity service composed of virtual APs over a set of physical Aps, and ii) Allocate a connection through the wireless backhaul, which transport the traffic from such access service until a fiber attachment point.

Technically, *SWAM* is composed of the following components: i) the physical radio nodes featuring multiple interfaces used for access and backhaul (wireless Network Interface Cards – NICs); ii) a software-based data-path running on each physical radio node; and iii) the *SWAM* controller, featuring a backhaul module to instantiate paths over the wireless BH, a provisioning module, used to manage the lifecycle of virtual Access Points (vaps), and an access bridge module used to connect the vaps to the connections in the wireless backhaul.

The core of *SWAM* is the software based datapath depicted in the right part of Figure 4-18, where we can see an example of a node with three physical wireless interfaces and one Ethernet interface. One wireless interface is used to serve access traffic and instantiates two vap interfaces for tenants A and B, whereas the other two wireless interfaces are used for wireless backhaul and instantiate two backhaul (mesh) interfaces. The Ethernet interface connects to the wired network and instantiates a tunnel interface.

The goal of the *SWAM* datapath is to process packets coming from the tenants’ customers (vap interfaces) and deliver them to the appropriate *SWAM* gateways through the wireless backhaul (mesh interfaces). A three level hierarchy of software switches is used for this purpose: i) Per-tenant access bridges, ii) the integration bridge (br_int), and iii) the backhaul bridge (br_bh). The core idea behind the *SWAM* datapath is a logical separation between the access and the backhaul (BH). The job of the wireless BH is to forward packets along a set of end-to-end tunnels, whereas the job of the access side is to match traffic coming from the tenants’ vaps to the appropriate BH tunnels. In *SWAM*, a BH tunnel is defined using a VLAN tag, and provides a unidirectional connection between two interfaces of a per-tenant access bridge. A detailed view of the *SWAM* datapath is depicted in the right part of Figure 4-17, and the interested reader is referred to [4-43] for a detailed description.

To evaluate the performance of *SWAM*, the right part of Figure 4-17 shows the CDFs of the overall handover and BH tunnel reallocation time measured for two different devices in an indoor testbed, which on average is around 30ms. The left side of Figure 4-17 depicts an uninterrupted TCP session when a *SWAM* gateway is reallocated by the controller for load balancing purposes.

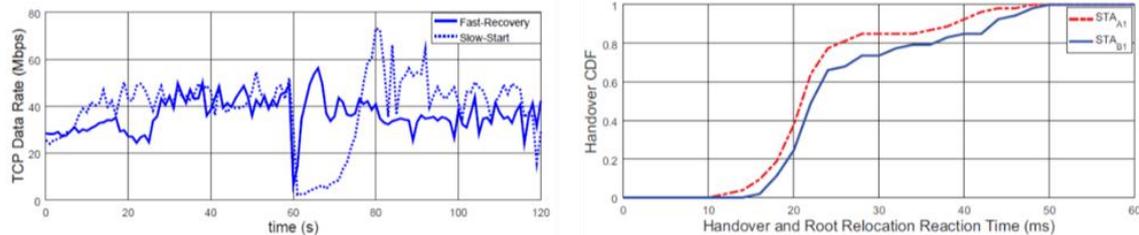


Figure 4-17 SWAM Mobility Evaluation

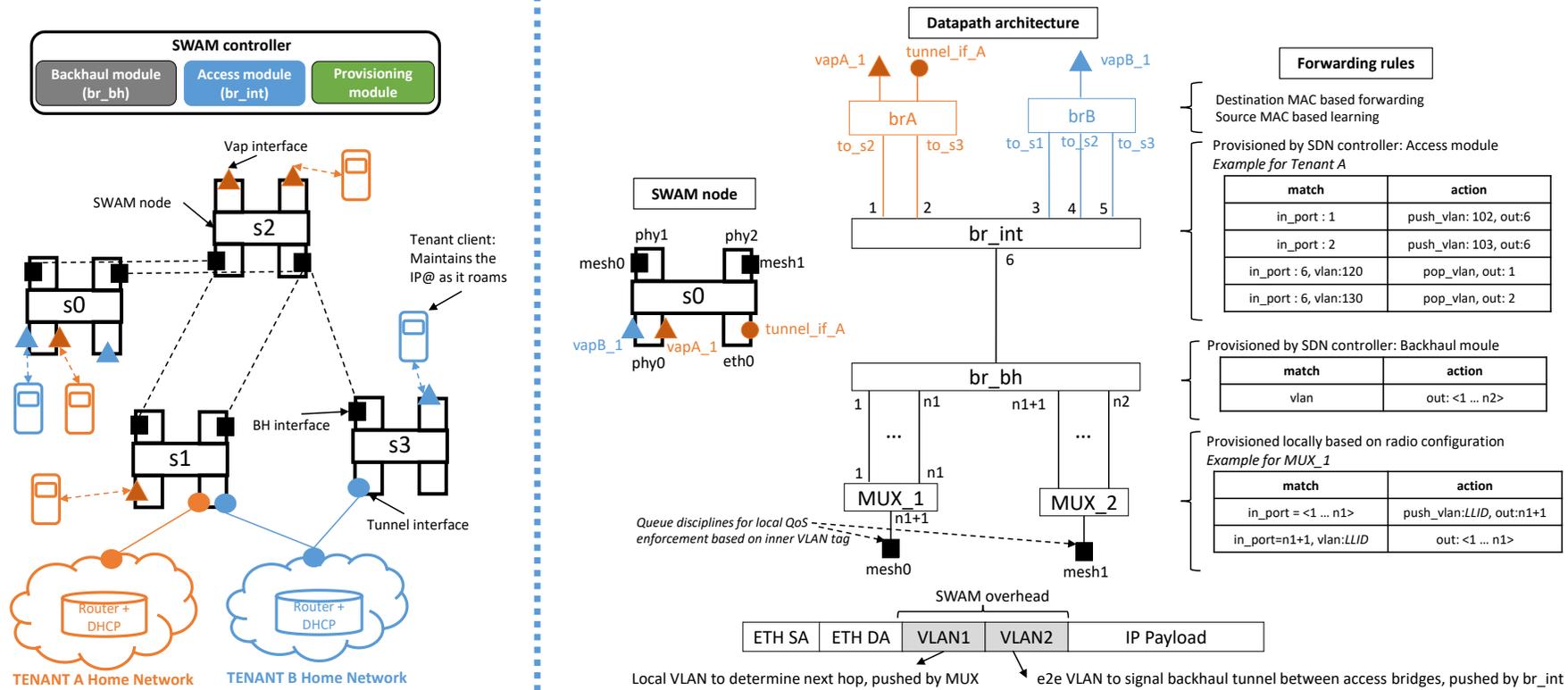


Figure 4-18. Design of SWAM

4.3.2.3 Satellite Backhaul

A satellite system can be used as a transport network within the 5G network in order to provide connectivity between areas. The BH between the AN and the CN can therefore rely on such system. Satellite systems remain the only or the most viable system to provide connectivity in specific contexts. Satellite and terrestrial integration in 5G can be investigated around four main uses cases that are identified in [4-44]:

- **Use Case 1: “Edge delivery & offload for multimedia content and MEC VNF software”**: Providing efficient multicast/broadcast delivery to network edges for content such as live broadcasts, group communications, Multi Access Computing, VNF update distribution
- **Use Case 2: “5G fixed backhaul”**: Broadband connectivity where it is difficult or not (yet) possible to deploy terrestrial connections to towers (remote/isolated areas);
- **Use Case 3: “5G to premises”**: Connectivity complementing terrestrial networks, such as broadband connectivity to home/office small cell in underserved areas in combination with terrestrial wireless or wireline;
- **Use Case 4: “5G moving platform backhaul”**: Broadband connectivity to platforms on the move, such as airplanes, trains, or vessels.

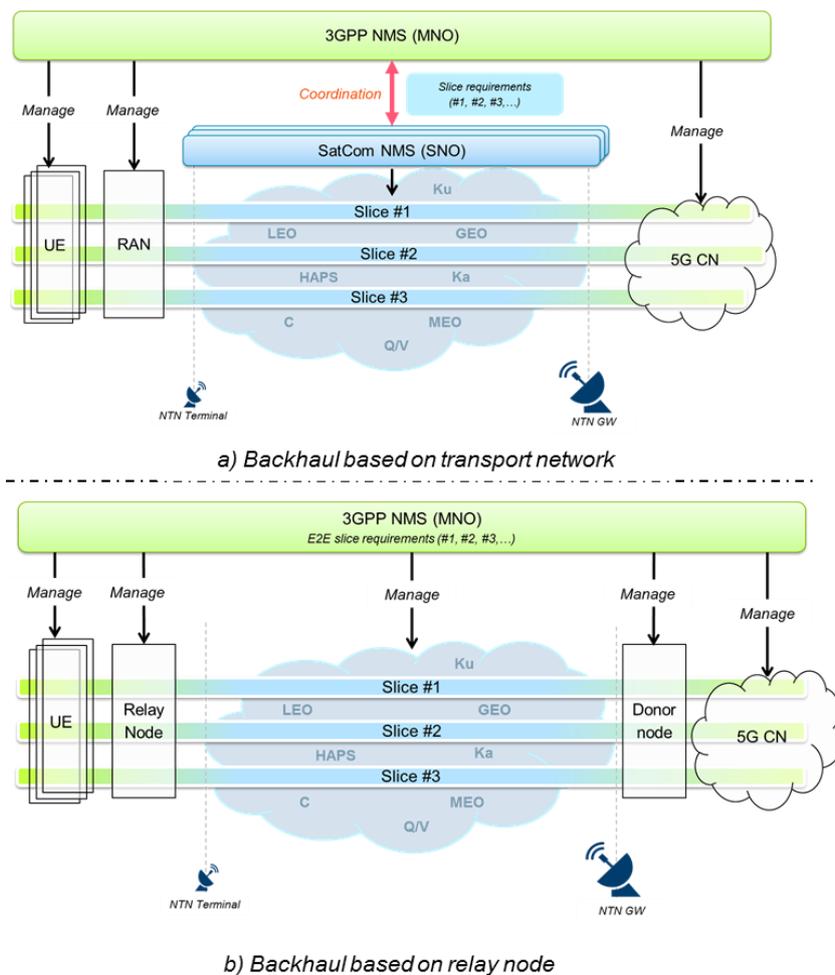


Figure 4-19 Satellite backhaul implementation options

The identified indirect access implementation options can be classified in two main categories, as proposed in [4-45] and [4-46].

- **Transport Network (TN)**, depicted in Figure 4-19 a), where the satellite network offers transport features to the 5G network between the 5G core and the RAN. The TN interfaces provide enhanced management and advanced satellite network functionalities (e.g. 5G QoS adaption to satellite class of service, dynamic satellite resources management, etc.). The backhaul implementation based on TN includes two implementation options, mainly differentiated by the features provided by satellite network at the interfaces with the terrestrial network. These interfaces can be natively 5G ready (TN based on 3GPP system specifications) or would require a development of an adaptation layer (TN not based on 3GPP system specifications);
- **Relay Node based implementation options (RN)**, depicted in Figure 4-19 b), representing a satellite-capable UE endorsing a relay functionality (i.e. multiplexer node role) which can serve other UEs and being backhauled to the ‘donor RAN’ and 5G CN through a satellite link. This approach includes three implementations options, differentiated by the type of access between the RN and the 5G CN: 3GPP access, trusted non-3GPP access and untrusted non-3GPP access.

All backhaul implementation options need to support **Multi-access Edge Computing (MEC)** as a key 5G feature. This includes edge delivery and support of network function (NF) delocalisation. In case of multilink support (satellite and non-satellite links), Hybrid Multiplay Functions are foreseen in order to improve service Quality of Experience (QoE). Traffic steering, switching and splitting would therefore being performed at (R)AN level over different available backhaul links. All these challenges are summarized in Table 4-2.

Table 4-2 Implementation options and key challenges for direct and indirect access

Backhaul Implementation option	Key challenges	Network management	Potential additional supported features
Relay node with 3GPP Access	<ul style="list-style-type: none"> • NR over satellite • Adaptation of relay node mechanisms to satellite terminal 	Single integrated NMS	<ul style="list-style-type: none"> • Edge delivery • NF delocalisation • Hybrid multiplay (traffic steering at RAN level) • Enhanced UP, CP, MP interfaces between satellite domain and terrestrial domain
Relay node with Trusted non-3GPP Access	<ul style="list-style-type: none"> • Make satellite access a trusted non-3GPP access in standards • Adaptation of relay node mechanisms to satellite terminal 		
Relay node with Untrusted non-3GPP Access	<ul style="list-style-type: none"> • Implement untrusted access mechanisms as requested by 5G standards • Adaptation of relay node mechanisms to satellite terminal 		
Transport Network based on 3GPP System specification	<ul style="list-style-type: none"> • Design a specific “5G ready” satellite transport network based on 5G system specifications 	3GPP NMS & Sat NMS working in coordination	
Transport Network not based on 3GPP System specification	<ul style="list-style-type: none"> • Design an adaptation layer for existing satellite transport network 		

For efficient 5G satellite and terrestrial integration, the support of network slicing by all the domains is a key requirement. SDN/NFV paradigms applied to satellite communications have been identified as key assets to provide appropriate tools and interfaces in order to ensure efficient support of end-to-end network slicing.

Management approaches of the future integrated satellite-terrestrial 5G network have been analysed and the two main approaches regarding the Network Management System (NMS) are:

- **Separated NMSs with coordination between the 3GPP NMS and the satellite NMS:** in this case, the 3GPP NMS only manages the terrestrial 3GPP components, while the satellite components are entirely managed by a separate management system (satellite NMS). Coordination between the two NMSs is therefore foreseen for an efficient resource usage and to ensure appropriate responses to the requests (e.g. service, monitoring, etc.) from one domain to another. This approach is typically applicable to backhaul implementation option based on satellite transport network (see Figure 4-19, a);
- **Single integrated network management:** in this case, the 3GPP NMS ensures the management of the whole satellite-terrestrial network, including the satellite terminal. This approach is typically foreseen for relay node implementation cases in which the satellite terminal acting as a relay node would be managed by the same entity managing the terrestrial network i.e. the 3GPP NMS (see Figure 4-19, b).

4.3.3 Fiber-Wireless (FiWi) PtMP

A point-to multipoint (PtMP) network that interconnects eCPRI capable equipment is proposed in [4-47]. This solution exploits:

- i) the high spectral efficiency of analog RoF transmission,
- ii) the large spectrum of the V-band, and
- iii) the audacity of OBFNs and massive MIMO antennas.

This way it solves the problem of cell densification since it allows for flexible wireless last-mile placement of the Remote Radio Heads (RRHs) in the area of service while maintaining compatibility with standardized eCPRI equipment. The architecture is versatile and supports operation as BH, MH, and FH depending on the placement of the technology in the 3GPP Centralized Unit (CU)/Distributed Unit (DU)/ Remote Unit (RU) stack. Figure 4-20 depicts the implementation of the three abovementioned operation modes.

The solution represents a Fiber-Wireless (FiWi) PtMP bridge between the CU and the RU(s), linking the multiple Service Access Points (either gNB, DUs or RUs depending on whether the network is used for BH, MH or FH).

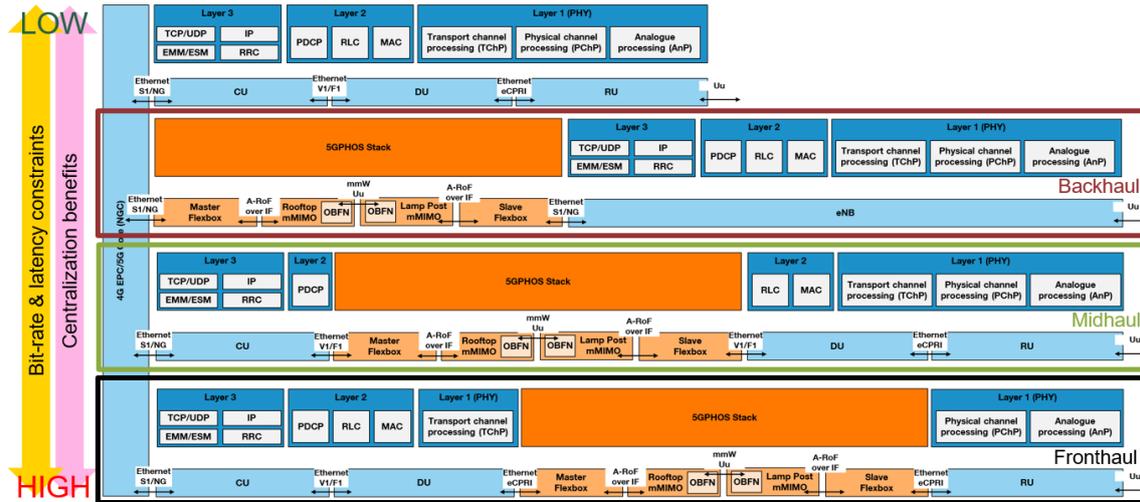


Figure 4-20 PtMP solution for the transport network: BH, MH, and FH implementations

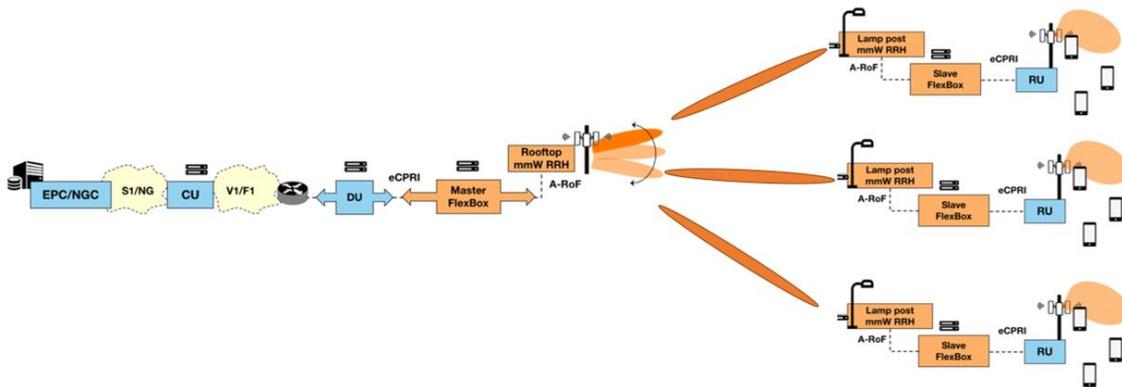


Figure 4-21 Example of a PtMP FH solution used to wirelessly bridge multiple RUs with a DU

An example of the solution when used in FH operation is depicted in Figure 4-21, and involves several units, which leverage ARoF transmission to a Rooftop RRH which, in turn, has the ability to steer the transmission using an OBFN to different lamp posts at mmWave frequencies (V-Band) employing multiple sub-bands and using pencil beams (Figure 4-22 a).

The eCPRI traffic represents the Data Plane (DP), whereas the C&M and Synch traffic together make up the Control Plane (CP). The eCPRI traffic spans three main categories:

- eCPRI traffic: the actual eCPRI traffic that contains the user data, the real time control as well as the rest packets required for the services. The eCPRI traffic is passed to the Ethernet MAC layer through UDP and IP layers.
- Control and management data (C&M): this data carries the control and management traffic that goes through to the remote unit. This data is carried through management protocols, such as SNMP, over UDP/TCP and IP layers.
- Synchronization data (Synch): Data employed in order to synchronize the clocks of the remote and centralized units. It uses primarily the PTP protocol (running on top of UDP and IP layers) but can be a combination of PTP with synchronous Ethernet to achieve the highest possible accuracy.

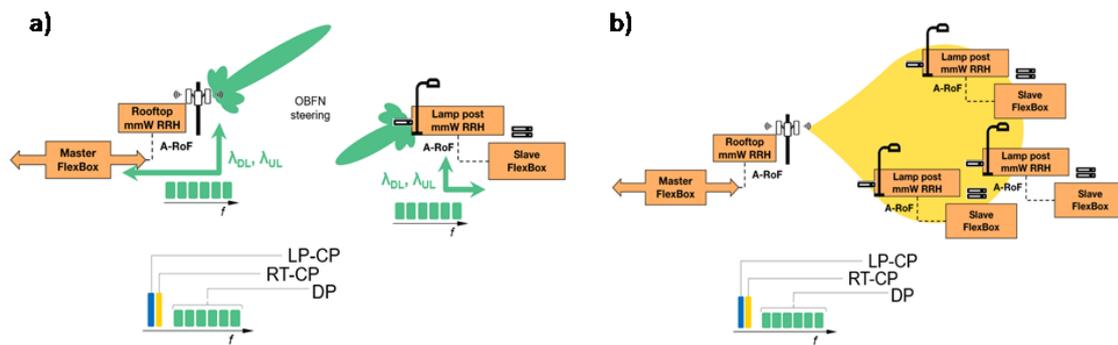


Figure 4-22 (a) Data Plane pencil beams (b) The Control Plane wide beam antennas

The control plane employs out-of-band (meaning separate than the DP frequencies) channels to distribute the C&M and sync messages to/from the lamppost RRHs. These channels are called the Control Plane Lamppost channels (CP-LP). In the downlink direction there is also one extra channel used for control of the Rooftop antenna (the CP-RT). The CP-LP is broadcasted to the lampposts through a wide beam that covers all the served lamppost antennas (Figure 4-22 (b)). In the DL direction there is no challenge since it is only the rooftop that transmits. In the UL direction however, we have to deal with the multiple access problem. This problem can be solved in two ways, based on the decisions of the infrastructure owner:

1. The Frequency Division solution (FDD): In this approach, we assign one UL CP-LP band to each lamppost. Since the traffic in the CP-LP is very low, these bands need to have only a small bandwidth. A variant could be to use an Orthogonal Frequency Division Multiplexing (OFDM) approach. In OFDM, multiple access is achieved by assigning subsets of subcarriers to individual users. This allows simultaneous low-data-rate transmission from several users.
2. The Code Division Multiple Access solution: In this approach several transmitters can send information simultaneously over a single communication channel by employing spread spectrum technology and a special coding scheme where each transmitter is assigned a code.

In this architecture, flexibility in the resource allocation comes at in two forms:

1. In the case of employing a single wavelength, the resource allocation method can assign the DP sub-bands to the lampposts depending on their traffic demands. For instance, considering a high layer split where the traffic in the FH fluctuates depending on the actual user traffic, one lamppost can receive more sub-bands for communicating with the Rooftop antenna, whereas others receive only one or their DP can be shut down completely in lack of users.
2. In the scenario, where we employ WDM and therefore multiple wavelengths in the optical domain, the flexibility comes from turning on or off various Rooftop antennas, so as to distribute the traffic stemming from the lampposts to more rooftops and essentially to other wavelengths

This flexible solution allows for RRH/Small Cell densification without the need for new fiber installation to all lampposts. Using this solution, an infrastructure owner could provide the framework to operators to install their RRHs/Small Cells to the lamp-posts antennas, so as to provide coverage in the specific area. Two or more RRHs/Small Cells can be connected to the same lamppost antenna and therefore share the same sub-bands. Another way to multi-tenancy stems from the installation of different functional-split RRHs, meaning that each operator is free to make its choice of desired functional splits.

4.4 Examples of data plane programmability

4.4.1 Programmable Data Plane

A programmable Data Plane has been designed and prototyped in the context of 5G PPP, which can be deployed as the data path in non-RAN segments such as the Edge Network, the Transport Network and the CN [4-8]. The purpose is to enable network traffic/slice Quality of Service (QoS) control in the data plane, and thus enable QoS-aware network slicing.

Figure 4-23 shows an overview of the programmable data plane architecture. In this simplified overview diagram, two host machines are illustrated as an example representation of the Edge Network (MEC) and the Core Network (CN) respectively. The Edge and the Core Networks are interconnected through the Transport Network. The antenna and the DU are connected to the MEC compute via a physical switch. The solid red circles indicate possible programmable points in the data path for traffic control, etc.

This architecture supports hardware-based, software-based or hybrid data plane programmability. The diagram shows a hybrid approach, combining both software- and HW-based schemes. The HW-based scheme leverages the programmability at the HW, especially the network interface cards (NICs), whilst the software-based scheme explores software data paths such as Open vSwitch (OVS), Data Plane Development Kit (DPDK) as well as virtual NICs. In terms of performance, the hardware-based approach is recommended; meanwhile, the software-based or hybrid approach would provide a more cost-efficient and flexible solution. Therefore, the choice of a specific approach depends on specific use cases.

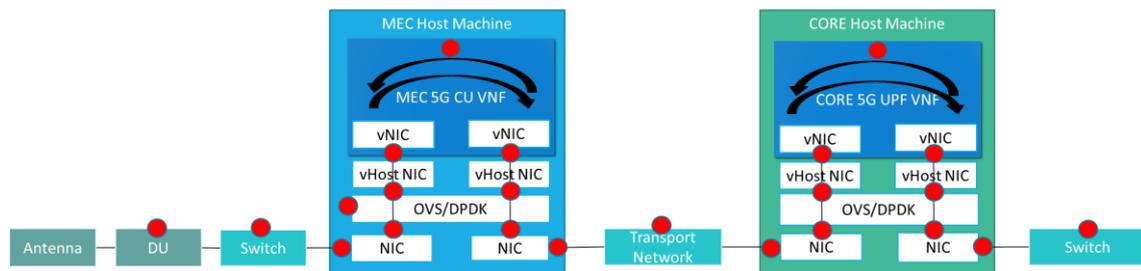


Figure 4-23 Programmable data plane architecture [4-8]

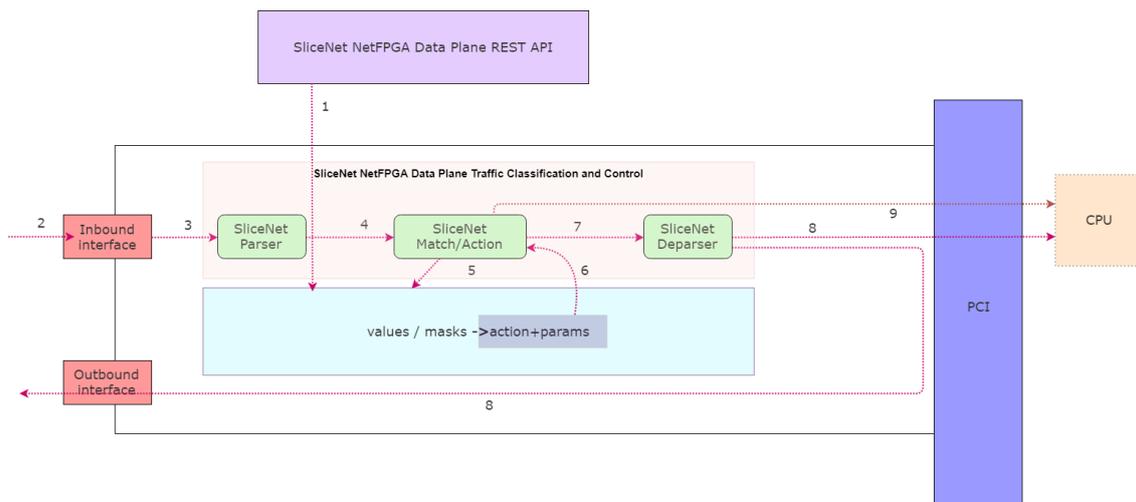


Figure 4-24 Programmable data plane prototype (HW-based)

For prototyping purposes, the hardware-based approach has been implemented to boost the performance at the data plane. The target is to allow traffic control and thus provide QoS-aware

network slicing at the data plane, whilst minimising the overhead occurred especially the extra delay introduced. Figure 4-24 shows the architecture of this prototype, based on the P4 NetFPGA [4-11] platform. To be brief, the workflow follows that of the SimpleSumeSwitch [4-12]: traffic flows go through the Parser, Match/Action and Deparser pipeline to be classified and processed accordingly, based on the definitions of the different network slices in the data plane. Specific traffic control actions regarding a flow or flows belonging to a network slice can include dynamically configuring priorities, dropping, mirroring to another interface, or being further processed by the CPU, and so on.

4.4.2 Stateful Packet Processing in Hardware

In recent years, the current solutions in terms of data plane programmability present several shortcomings that prevent the nowadays-available programmable data planes to act directly for stateful functionalities. One of the bigger limitations in the current programmable data planes is the absence of a clear per-flow stateful model for storing directly in the data plane the information gathered on the different flows under analysis. The goal of executing stateful network functions (NFs) to enhance the programmability of transport solutions could not yet be accomplished.

The HW constraints in terms of memory amount and number of operations that can be executed for each packet that must be processed by the network, pose severe limitations to the programmable data plane architecture. These limitations are in contrast to the requirements in terms of flexibility of the NFs. Fortunately, there is some recent research work showing that the main hardware elements composing the data plane could provide enough flexibility and programmability to realize several network functionalities directly in the data plane [4-29].

Recently, examples of programmable data planes [4-30] emerged as ideal target devices to implement these complex NFs without the intervention of the control plane. These programmable data planes can be configured using specific programming languages (such as P4 or POF), and will be able to provide protocol independence, thus managing programmable parsing of the protocol stack for generic field extraction and packet encapsulation/decapsulation. These solutions are being extended both in terms of switch matching capabilities and in terms of actions to apply to the processed packets using programmable pipelines of match/action stages.

The envisioned architecture of FlowBlaze is presented in Figure 4-25 [4-31]. This architecture is conceived to supersede the above mentioned limitations of programmable data plane retaining the ability to sustain wire speed packet processing.

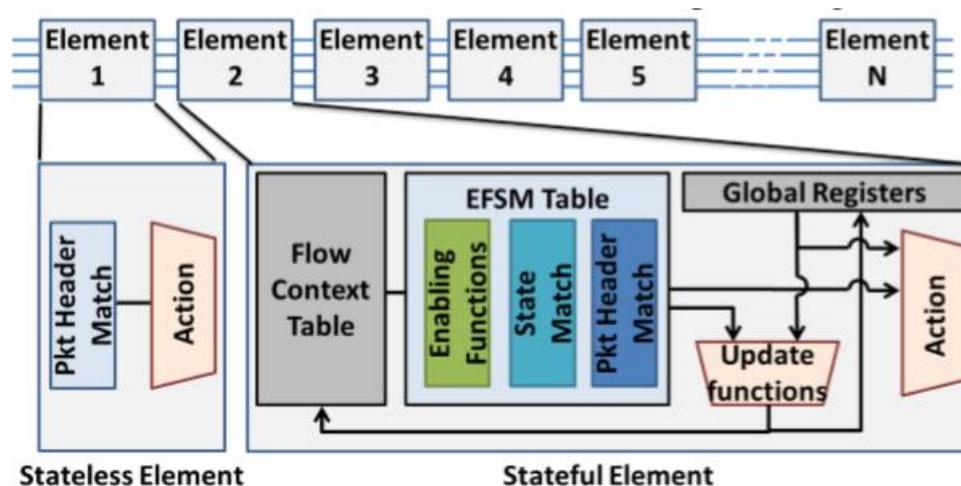


Figure 4-25 Initial architecture design of the stateful packet processing [4-32]

FlowBlaze is composed by a pipeline of elements, where each element can be a stateless element (an OpenFlow/P4 match-action table) or a stateful element, able to execute per-flow eXtended Finite State Machine (XFSM), thus providing stateful functionalities. There are two types of stateful elements in programmable data planes: tables and counters/registers. Nowadays tables can be controller only from the control plane (insertion/update/delete operation can be executed only using specific control-plane commands). Registers/counters array can be updated directly in the data plane, but it is hard to map a row of the array to a specific flow. The mapping between the flow and the array elements can be also realized using a matching table, but the use of this approach prevents the data plane update of the table (e.g. when the arrival of a new flow or its expiration require the involvement of the control plane). FlowBlaze solves these issues providing a specific type of table that can be updatable directly in the data plane. The design of an efficient data plane updatable table while retaining wire speed is a challenging engineering task, as discussed in deliverable 3.1 [20]. From the programmability point of view this enables the in-data plane management of several per-flow network functions, spanning from configurable Network Address Translation (NAT) services to flow monitoring, from QoS policies up to the deployment of data-driven routing/forwarding mechanisms.

The FlowBlaze technology is an enabler for flexible functional splits. The data plane level stateful per-flow functionalities permit to avoid the latency overhead and throughput bottleneck for the network function primitives used to provide the functional split. An example of network functionalities on top of FlowBlaze is the development of a routing algorithm for DCs able to dynamically estimate the best path in terms of latency/throughput. This will allow to take routing decisions depending on the flow requirements, i.e. forwarding latency critical flows using the low latency paths.

Finally, it is true that the resource disaggregation concept allows efficient provisioning of the HW and SW resources that are available in the network. However, the actual use of these resources must feature some functionalities (that can be at least roughly identified as network functions) to be independently deployed and executed in different heterogeneous computing resources. In principle, the same function could be realized in a fixed functionality ASIC chip in a highly specialised processor (DSP for signal processing or in a network processor for packet level operations), in an FPGA, in a GPU or in an off the shelf x86 host. All these resources have different programming languages and different interfaces with the external world. The obvious solution for designing the same functions in all possible platforms on which the function can be executed is not scalable. This is due to the large number of platforms and to the very different programming models that need to be applied. To solve this issue, an ad-hoc high level programming language (called XL, the XFSM Language) is able to describe the per-flow network functions which will be executed by a FlowBlaze engine, regardless of its specific hardware implementation (ASIC,FPGA or SW).

4.4.3 Segment routing

To date, the way to slice the transport network in the MAC and the IP layers was relying on MPLS label tagging and VPN tunnelling, where each flow was identified with a specific label and/or transferred within a specific VPN tunnel, respectively. For each slice/flow, bandwidth was guaranteed in SW, through packet classification and traffic rate limitation on each hop. In the light of network slicing, VPN enhancements have recently appeared in order to provide dedicated network resources for each network slice, based on the slicing capability of the network infrastructure (e.g., nodes, links) and integration between overlay and underlay networks. Key requirements are guaranteed performance and isolation between different network slices and sharing when possible for services within the same slice. However, these aforementioned solutions are not able to meet other objectives like, for example, delay guarantees or fast routing protocol convergence times. Segment Routing (SR) is seen as an exploitable technology that is

able to provide service guarantees and support advanced functionalities for the virtualised network in Layer 3.

SR is a new protocol [4-34] [4-35] designed to forward data packets on a network based on the source-based routing paradigm. It is expected to play a key role in deterministic networks and networks where “plain” VPN solutions are not enough, since besides encryption by means of performance existing VPN solutions are actually best effort and are not able to provide service guarantees to the virtual “sliced” network. Even when MPLS-TE solutions are deployed the end-to-end network performance is subject to the routing protocol behaviour and the policy used.

An example architecture based on the source routing paradigm [4-36] seeks the right balance between distributed intelligence and centralised programmability. Instead of performing routing based on a node to node basis, SR divides a network path into several segments. Each forwarding path is constructed based on sequentially arranged segment list. A segment may be associated with a service instruction, with a node, a link or a path.

SR achieves source routing by steering packet through a list of segments (SIDs), where SIDs are used to represent topological, service or other instructions, and it leverages the source routing paradigm. An ingress node steers a packet through an ordered list of instructions (i.e. segments). Each one of these instructions represents a function to be called at a specific location in the network. A function is locally defined on the node where it is executed and may range from simply moving forward in the segment list to any complex user-defined behaviour. Network programming consists of combining segment routing functions, both simple and complex, to achieve a networking objective that goes beyond mere packet routing.

SR offers a number of benefits like simplification of the control plane of MPLS type of networks, efficient topology independent-loop-free alternate fast re-routing protection, higher network capacity expansion capabilities, smooth integration of SDN technology while it can also be used as an enabling technology for deterministic networking.

4.5 References

- [4-1] 3GPP TS 23.501, “System Architecture for the 5G System”.
- [4-2] 3GPP TS 29.500, “Technical Realization of Service Based Architecture; Stage 3”.
- [4-3] Hypertext Transfer Protocol (HTTP) over multicast QUIC, <https://datatracker.ietf.org/doc/draft-pardue-quic-http-mcast/>
- [4-4] 3GPP TS 29.244, “Interface between the Control Plane and the User Plane Nodes; Stage 3”.
- [4-5] 3GPP TS 38.413, “NG Application Protocol (NGAP)”.
- [4-6] 3GPP TS 38.412, “NG signalling transport”.
- [4-7] D. Camps-Mur et al., “5G-XHaul: A novel wireless-optical SDN transport network to support joint 5G backhaul and fronthaul services”, to appear in IEEE Communications Magazine.
- [4-8] SliceNet Deliverable D3.1 “Design and Prototyping of SliceNet Virtualised Mobile Edge Computing Infrastructure”, https://doi.org/10.18153/SLIC-761913-D3_1
- [4-9] 3GPP TS 26.346, “Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs”.
- [4-10] 3GPP TS 29.116, “Representational state transfer over xMB reference point between Content Provider and BM-SC”.
- [4-11] The P4 Language Consortium, “P4→NetFPGA: A low-cost solution for testing P4 programs in hardware.” Feb 2018. [Online]. Available: <https://p4.org/p4/p4-netfpga-a-low-cost-solution-for-testing-p4-programs-in-hardware.html>.

- [4-12] SimpleSumeSwitch Architecture, Feb. 2018. [Online]. Available: <https://github.com/NetFPGA/P4-NetFPGA-public/wiki/Workflow-Overview>.
- [4-13] 5G-Xcast Deliverable D4.1: “Mobile Core Network”.
- [4-14] 5G-Xcast Deliverable D4.3: “Session Control and Management”.
- [4-15] RFC 3926, “FLUTE - File Delivery over Unidirectional Transport”.
- [4-16] 5G-Xcast Deliverable D2.4: “Analysis and Deployment of Terrestrial Broadcast in 5G-Xcast”
- [4-17] S. Yan et al., “Multilayer network analytics with SDN-based monitoring framework”, in IEEE Journal of Optical Communications and Networking, vol. 9, 2017, pp.A271-A279.
- [4-18] Tzanakaki et al., “5G infrastructure supporting end-user and operational services: The 5G-XHaul architectural perspective”, IEEE International Conference on Communications (ICC 2016), May 2016.
- [4-19] E. e. a. I. Hussain, GMPLS Routing and Signaling Framework for Flexible Ethernet (FlexE), IETF draft-izh-camp-flexe-fwk-02, 2016.
- [4-20] Huawei technologies, RH2288 V3 Server User Guide, Huawei technologies, 2017.
- [4-21] Huawei technologies, OptiX PTN 990 Product Documentation- Product Version : V100R008C10SPC500, Huawei Technologies, 2018.
- [4-22] Huawei technologies, OptiX PTN 990 Product Documentation- Product Version : V100R008C10SPC500, Huawei Technologies, 2018.
- [4-23] ‘Metro-Haul project website’, <https://metro-haul.eu>, accessed 14 April 2019
- [4-24] Farhadi Beldachi, et.al.: ' Experimental Demonstration of 5G Fronthaul and Backhaul Convergence based on FPGA-based Active Optical Transport', in Proc. OFC (2018).
- [4-25] ‘Facebook Voyager Optical Interfaces’, <https://docs.cumulusnetworks.com/display/DOCS/Facebook+Voyager+Optical+Interfaces>, accessed 14 April 2019.
- [4-26] BlueSPACE Deliverable D2.2 v1.0: “Physical architecture, system and network requirements”, June 2018.
- [4-27] Raul Muñoz et al. “BlueSPACE's SDN/NFV Architecture for 5G SDM/WDM-enabled Fronthaul with Edge Computing.” EUCNC 2018, Ljubljana, Slovenia, 18-21 June 2018.
- [4-28] ‘WaveShaper 1600A’, <https://www.finisar.com>, accessed 14 April 2019
- [4-29] R. Miao, et al. "SilkRoad: Making Stateful Layer-4 Load Balancing Fast and Cheap Using Switching ASICs," in Proc. ACM SIGCOMM, 2017.
- [4-30] P. Bosshart, et al. "Forwarding metamorphosis: Fast programmable match-action processing in hard-ware for SDN." ACM SIGCOMM Computer Communication Review. Vol. 43. No. 4. ACM, 2013.
- [4-31] 5G-PICTURE Deliverable D3.1, “Initial report on Data Plane Programmability and infrastructure components”.
- [4-32] 5G-PICTURE Deliverable D4.1, “State of the art and initial function design”, April 2018.
- [4-33] 5G-PICTURE Deliverable D3.2, “Intermediate report on Data Plane Programmability and infrastructure components”
- [4-34] IETF RFC 8402, “Segment Routing Architecture,” 2018.
- [4-35] K. M. K. T. Clarence Filsfils, Segment Routing, Part I, Cisco Systems, 2016.
- [4-36] C. Filsfils, S. Previdi, G. L., B. Decraene, S. Litkowski and R. Litkowski, “Segment Routing Architecture”, 2018.
- [4-37] 5G-PICTURE Deliverable D2.2, “System architecture and preliminary evaluations”, July 2018.
- [4-38] M. Ehrig, M. Petri, V. Sark, J. G. Teran and E. Grass, “Combined high resolution ranging and high data rate wireless communication system in the 60 GHz band”, 11th Workshop on Positioning, Navigation and Communication (WPNC), pp. 16, 2014.

- [4-39] N. Maletic, V. Sark, J. Gutiérrez and E. Grass, "Device Localization Using mmWave Ranging with Sub-6-Assisted Angle of Arrival Estimation," 2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Valencia, 2018, pp. 1-6.
- [4-40] 5GENESIS Deliverable D2.2 "5GENESIS Overall Facility Design and Specifications", December 2018.
- [4-41] X. Song et al., "Design and Experimental Evaluation of Equalization Algorithms for Line-of-Sight Spatial Multiplexing at 60 GHz," in IEEE Journal on Selected Areas in Communications. doi: 10.1109/JSAC.2018.2872286.
- [4-42] M. Grandi, D. Camps-Mur, A. Betzler, J. J. Aleixendri, and M. Catalan, "SWAM: SDN-based Wi-Fi Small Cells with Joint Access-Backhaul and Multi-Tenant Capabilities", IEEE IWQoS 2018.
- [4-43] 5G-PICTURE Deliverable D4.2, "Complete design and initial evaluation of developed functions", November 2018, https://www.5g-picture-project.eu/download/5g-picture_D4.2.pdf.
- [4-44] SaT5G Consortium. Deliverable D2.1 "Satellite reference use cases and scenarios for eMBB". August 2017.
- [4-45] Sat5G Consortium. Deliverable D3.1, Integrated SaT5G General Network Architecture. November 2019.
- [4-46] Sat5G Consortium. Deliverable D3.2, Integrated SaT5G Detail Network Architecture. December 2018.
- [4-47] G. Kalfas et al., "Next Generation Fiber-Wireless Fronthaul for 5G mmWave Networks," in IEEE Communications Magazine, vol. 57, no. 3, pp. 138-144, March 2019.
- [4-48] Gartner, "Magic Quadrant for BI platforms. Analytics Value Escalator", 2012.
- [4-49] E. Pateromichelakis et al. "End-to-End Data Analytics Framework for 5G Architecture," in IEEE Access, 2019.

5 Management & Orchestration Architecture

5.1 Introduction

This chapter deals with the question of managing and orchestrating services running in 5G networks. This pertains to high-level architecture questions about how to structure this problem: how to describe functions (virtualized or physical), how to connect them into services, and how and where to deploy them. The high-level architecture for this is pretty much settled by now, but a closer inspection reveals that there are still a lot of design choices left open within the constraints of such a settled meta-architecture. These choices are described and put into perspective with typical roles. Afterwards, the impact of DevOps is analysed.

5.2 Starting point: Existing High-level MANO architectures

Recently, the question how to structure the management and operation of a 5G system has received considerable attention. This question arose from the push towards softwarizing networking infrastructure, hoping for shorter time-to-market for new features, reduced cost, greater flexibility, versatility, and the ability to leverage a telecom operator's infrastructure for new business models, improving their competitive relation to mere over-the-top players.

All these goals, however, necessitate a better ability to manage the lifecycle of these software components running inside a network – the old adage of “the network is the computer” is coming true. This lifecycle management issue has been addressed as the Management and Operations challenge, for which a reference architecture has already emerged. This architecture originated from work in the context of an ETSI working group and has received additional fine-tuning over the last few years, both from an ETSI context as well as from various 5G PPP projects.

This section summarizes developments in ETSI and 3GPP, deriving a consensus “meta architecture”.

5.2.1 ETSI summary

5.2.1.1 ETSI NFV & MANO

The ETSI Industry Specification Group (ISG) on Network Functions Virtualization [5-1] is perhaps the most influential of ETSI's recent efforts; it helped to kick off the network functions virtualization field on a large scale. ETSI defined basic notions (NFV, VNF, NFVI, architecture, key interfaces, etc.) that still are shaping the field to a large degree. Much of the remainder of this document is based on this early initiative.

5.2.1.2 ETSI ZSM

The Zero touch network and Service Management Industry Specification Group (ZSM ISG) in ETSI focuses on service automation and management that leverages the principles of NFV and SDN [5-2]. The goal of ZSM is to define a new, future-proof, E2E operable framework enabling agile, efficient and qualitative management and automation of emerging and future networks and services. In a nutshell, the aim of ZSM is to have all operational tasks, including delivery, deployment, configuration, assurance, and optimization, executed automatically.

ETSI ZSM aims to facilitate the coordination and cooperation between relevant standardization bodies and open source projects. Its value proposition is providing guidance to the implementation

of management interfaces as well as coordinating and giving directions to achieve automated end-to-end network and service management solutions and architecture. As part of the end-to-end solution, the necessary management architecture and interfaces to support the end-to-end zero touch network and service management in multi-vendor environment are being identified.

5.2.1.3 ETSI ENI

The purpose of the Experiential Networked Intelligence Industry Specification Group (ENI ISG) in ETSI is to define a Cognitive Network Management architecture that is based on the “observe-orient-decide-act” control model (a variant of the well-known MAPE-K concept) [5-3]. It uses AI (Artificial Intelligence) techniques and context-aware policies to adjust offered services based on changes in user needs, environmental conditions and business goals. The system is experiential, in that it learns from its operation and from decisions given to it by operators to improve its knowledge of how to act in the future. ETSI ENI designs reference architecture to enable the use of AI in network operation and management. The named ENI engine, interfaces with the existing network to enhance the AI capability of the network. Up to now, ENI has developed use cases, requirements, and a preliminary architecture and interfaces. The work of ENI has been planned until 2021.

5.2.1.4 ETSI MEC

Multi-Access Edge Computing (MEC) is one of the key concepts for fulfilling some of the requirements of vertical services and can be regarded as a refinement of a more general MANO/NFV concept [5-4]. MEC and its integration in an NFV context was studied in the ETSI MEC017 document [5-5] and a reference architecture is provided with the following key observations:

- The mobile edge platform is deployed as a VNF and therefore the procedures defined by ETSI NFV for these means are used;
- ETSI NFV MANO sees mobile edge applications as regular VNFs allowing for reuse of ETSI MANO functionality (with perhaps some extensions);
- The virtualization infrastructure is deployed as an NFVI and its virtualized resources are managed by the VIM. For this purpose, the procedures defined by ETSI NFV Infrastructure specifications, i.e. ETSI NFV INF 003 [5-6], ETSI NFV INF 004 [5-7], and ETSI NFV INF 005 [5-8] can be used.

5.2.2 3GPP

3GPP-related activities relevant here are mostly SA2 (architecture) and SA5 (telecom management). SA2 heavily emphasises “network slicing” as a key concept in the core architecture (TS 23.501 [5-10]). A slice is seen as a logical network accessible to user equipment (UEs), extending across access & core for both user and data plane. A slice instance is seen as a set of network functions (similar term between ETSI and 3GPP) plus required resources; questions like identity of slices under dynamic function updates are still not fully settled.

SA5’s management perspective more directly relates to orchestration as discussed here. In particular, TR 28.801 more specifically talks about the relationship of services and slices and how to manage them [5-11]. But as slicing is not the focus of this chapter, we will ignore this discussion here.

Considerable effort had to be spent and is still necessary to put the ETSI and 3GPP views into perspective (EVE012 report [5-9]).

5.2.3 Status and consensus architecture

Based on these developments, a common structure is emerging for something that could be called a consensus architecture (or perhaps better called a consensus meta architecture, as it is not necessarily precise enough to be directly implementable). It comprises:

1. The control of individual network functions (the distribution of their software artefacts, the deployment on an execution environment, state management between instances of a function, bringing up or tearing down instances and deciding where to run how many instances of a function);
2. The chaining of individual functions into services (chains or general graphs), facilitated by different networking mechanisms (including, but not limited to software-defined networking);
3. The ability to use different underlying execution environments, ranging from different virtualization techniques (like virtual machines, containers, or even just plain processes) in clusters of vastly different sizes (from a simple additional CPU board in a base station to an entire large-scale data centre) over different, specialized, accelerated hardware (like FPGAs) to different networking environments (wireless, optics, cable) – sometimes this is called “technological domains”;
4. The ability to work with or across different administrative domains, encompassing different network operators (to provide a service at vast geographic ranges across multiple operators) or companies from different business models (e.g., network operators and separated cloud infrastructure operators); this is sometimes called “organizational domains”. In this context, it is worthwhile to emphasize that this discussion pertains to business roles but not necessarily to company organization – the same company can assume multiple roles or sometimes a single role can even be split across multiple companies (e.g. by subcontracting);
5. The ability to support a vast range of different applications with very different resource, deployment and orchestration needs as well as optimization goals (e.g., cost versus latency). This is sometimes called “application domains” (although this term is less well established than the previous ones and carries more connotations);
6. The idea to subdivide the infrastructure necessary to execute a service and carry its data in separate logical infrastructures with dedicated resources (or at least, guaranteed service performance) – commonly referred to as “slicing” – can be also be seen as part of a Management and Orchestration system; however, here the consensus is less clearly established than in the other areas. It is also conceivable to position a slicing system underneath or above a MANO system as well as inside it as an integral part.

Based on these six structural aspects, a number of core roles have emerged. These are described in more detail in the glossary, but essentially, we differentiate between:

- End user,
- Function developer,
- Application developer,
- Validation and verification entity,
- Tenant (owner of applications),
- Operator; typically, but not necessarily encompassing slicing operator; could also be separate
- Infrastructure provider; often divided further into network infrastructure provider, cloud infrastructure provider, etc.

Typical overlaps exist, from an application perspective, between function and application developer, validation entity and tenant, or application developer, validation, and tenant. From an infrastructure perspective, a typical conjoint role would be operator and infrastructure provider.

A less common, but certainly feasible example could be an operator acting as a tenant and running its own applications.

It is worthwhile to point out that these roles pertain to different phases in a service's lifecycle. They encompass the development of individual functions or entire services to their validation by possibly external, neutral entities, to the actual deployment and operation of these services. The interaction of these aspects – popularized in the software industry under the term “DevOps” – is an area that is still being explored by various projects and has not yet achieved full consensus.

A last aspect to point out is that some operations take place at very different time scales. This gives rise to a separation of “**orchestration**” actions (on long time scales, relatively heavy-weight operations like optimizing overall structure of a service or a group of services, perhaps also of slices) and “**control**” actions (on short time scales, relatively light-weight operations, e.g., the routing of a particular flow to a particular service instance). In some architectures (considered as refinements of the meta architecture described here), this leads to a separation of orchestration and control; however, this is not necessarily the case and not present in all considered architectures.

In the end, the typical components at a high level are still those defined by ETSI NFV: An NFV service platform is composed of an NFV orchestrator (NFVO), dedicated virtual network function manager (VNFM) and possibly physical network function (PNF) managers (PNFMs). A virtualised infrastructure manager (VIM) abstracts away details of how to manage deployment units (e.g., virtual machines vs. containers); VIMs can exist in many places and might then be called Edge Computing or Mobile Edge Computing (with little impact on the architecture as such). Concrete network configuration tasks (e.g., providing specific layer 2 connectivity) are typically outsourced to a separate SDN controller, working on behalf of the NFVO. Sometimes, the NFVO is split into two parts – orchestrator and controller as described above (akin to, but not identical with an SDN controller). Whether or not slices are supported and with what flavour is a strong differentiation between different architectures; sometimes, slice management is incorporated directly into the NFVO (with the argument that a network slice instance is nothing but a network service operating on guaranteed resources); sometimes, a separate slice manager is foreseen (with both NFVO triggering the slice manager and the other way around being options under consideration). Service management is sometimes separated from resource management, sometimes seen as an integral activity. In practically all cases, NFVOs can federate in some form with peer NFVOs, being in a single or in multiple organizations; sometimes, there is also a hierarchy of service management instances (starting from a more abstract multi-domain to specific single-domain MANOs in addition to – typically intra-domain – peered MANOs). At the lower layers, existing open source MANO platforms, including ONAP [5-23], OSM [5-24], and SONATA [5-25], are often used. Adaptors allow the mapping from abstracted activities towards more specific underlying implementation technologies.

Figure 5-1 summarizes these options by showing the most encompassing cases – it should be emphasized again that not in all concrete realizations of this meta architecture, all components or interfaces are indeed present. The left side shows a single-domain case, emphasizing relationship of SDN and MANO controllers towards actual resources, possibly abstracted away by a VIM. The right side shows a multi-domain case (simplifying resource aspects), highlighting possible relationships between multi-domain and single-domain service management (hierarchy vs. federation). An additional variant could also foresee orchestration functionality on top of the single-domain service management functions (see Section 5.3.1.2).

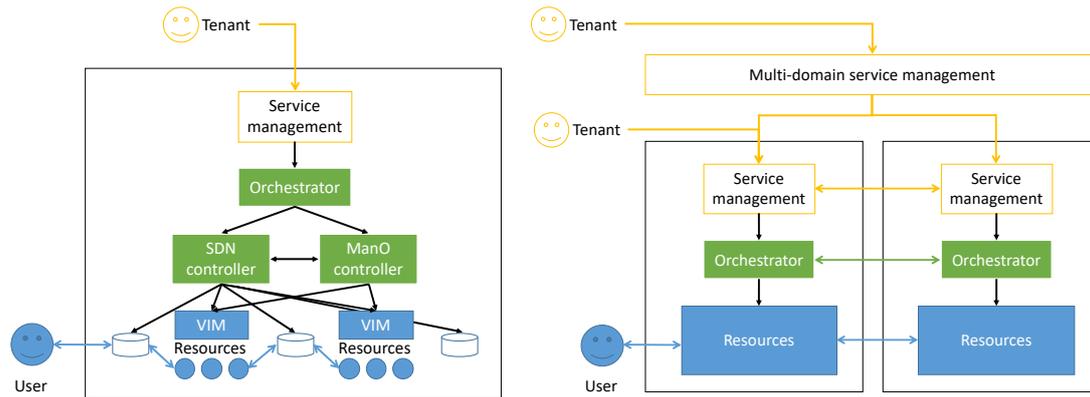


Figure 5-1 Meta-architecture for single-domain case (left) and multi-domain case (right)

5.2.3.1 Example aspect: Heterogeneity

To illustrate the versatility of this meta architecture, let us consider one example aspect: heterogeneity. This arises in several contexts. First, the MANO framework in its entirety needs to deal with vastly heterogeneous services with very different requirements, for example, on latency. A good MANO framework needs to be able to process such formalized requirements and deploy service functions close to the edge in order to keep latency low. We illustrate that in the following subsection by two example services (a push-to-talk and a mission-critical chat/content-delivery service). Second, heterogeneity also pertains to the underlying platform, both in terms of software and hardware infrastructure. These aspects are illustrated in the subsections afterwards.

5.2.3.1.1 Heterogeneous service deployment

An example where services with heterogeneous requirements have to be deployed are push-to-talk services or content-delivery services for mission-critical public-safety applications, coexisting with ordinary services with substantially lower service requirements [5-12]. Unless one subscribes to the notion of slicing with perfect control over all resources, there is a need for thought in the orchestration decisions.

A core insight is traffic locality: most traffic in such applications stays local, and inside the mobile edge. There is no need to involve the core network with data transport or function execution; in fact, keeping traffic and execution local improves roundtrip times and hence user satisfaction. Hence, an orchestrator needs to be aware of these local traffic properties, needs to have access to resources in the mobile edge, and then needs to deploy the corresponding network functions there and route traffic accordingly (this is of course orthogonal to any slicing aspects, but slices need to provide sufficient topology information to an orchestrator and need to exist at required places).

In detail, there are options how to share burdens between edge and core network. One option is to keep some administrative functions (e.g., call management) in the core and only move media-related functions to the edge. Alternatively, all relevant VNFs and services get moved to the edge. Trade-offs are operational complexity, the need to run multiple instances of the same services, reduced tunnelling overhead, and others. Similar options exist for chat and content delivery services in such mission-critical environments. Again, low latency here is the key objective.

5.2.3.1.2 Heterogeneous execution environments for network functions

A MANO system sits on top of an infrastructure. The job of the infrastructure is to provide actual resources (possibly via several mappings from virtual to virtual infrastructures before ending up on the actual, physical resources) to execute functions and to transport data between these functions. The infrastructure also provides an interface via which such function executions can be started, stopped, paused, moved elsewhere; the interface also provides means to influence the

transport of data (where the abstraction level of the first part of the infrastructure control interface is pretty much agreed upon; that of the second part is still debatable).

What this control interface does not, however, need to specify is the type of executing resources. On some abstraction level, it does not matter whether a given function is executed as a process, a container, a unikernel, a virtual machine, or in a hardware accelerator like an FPGA. What matters is that it *is* executing and that data can be transported in and out of it.

There are several possible separations of concerns with different trade-offs. Prominent options under discussion are as follows:

- Infrastructure hides its capabilities: There is no information flowing between infrastructure and MANO framework about which type of execution elements are available. It is the responsibility of the infrastructure management to choose the right realization of a function, where “right” entails both “functionally possible” (a VNF that only exists as a virtual machine image cannot be executed on a Container-only infrastructure) as well as “performance-optimal”.

While this is a convenient abstraction from the MANO system’s perspective, it does not seem feasible to realize. It assumes the existence of (essentially) every function in all possible execution forms and it surmises the infrastructure manager’s ability to decide what is “performance-optimal” – this does not seem plausible in the absence of information about the performance requirements of an entire service, and the relationships to other services as well.

- Alternatively, an infrastructure provides information about which type of execution resources are available, in which quantity, at which locations. The MANO framework can then use this as input to a corresponding optimization problem to choose the right combination of function executables and locations.

Apart from this decision/orchestration problem, there is also a typical control or “plumbing” problem. Making sure that data of a service chain flows correctly between functions sitting inside different virtualization systems (e.g., a Kubernetes cluster hosting some functions in containers, an OpenStack cluster hosting other functions inside virtual machines) is not trivial, but progress has been made to ensure that such heterogeneous service chains indeed work correctly.

With such heterogeneity support in place, new approaches become possible. For example, when using FaaS environments, the load adaptiveness of a service chain can be improved considerably as spinning up new instances happens much faster in such an environment. However, this is also a good example for the understanding of service/function semantics necessary inside the MANO framework.

5.2.3.1.3 Heterogeneous hardware

Incorporating hardware accelerators is a promising approach for many 5G networking functions, e.g., for signal processing in mobile base stations. There are some obvious challenges to do that, e.g., a network function should be available in multiple different formats, going beyond the virtual machine and container discussion of the previous example; it needs to be available, for example, in an FPGA implementation as well as a GPU implementation. This does raise some challenges for the description formats of functions and services and the binary formats as well, but this is relatively straightforward to handle.

What is less obvious to handle is that such multiple versions of a given function need to be orchestrated in different fashions. A particular challenge is the fact that, unlike CPUs, hardware accelerators can practically not be multiplexed unless such multiplexing is able to state-share between multiple functions that are prepared accordingly. Also, the deployment times on an FPGA can be very different from starting a container on a CPU (depending on previous state,

whether the whole or only parts of an FPGA need to be reprogrammed). Hence, they are fully dedicated to a single function. This leads to different algorithmic tradeoffs when planning which function to run on such accelerators.

5.2.3.1.4 Heterogeneous services

The discussion of separating services into “network services” (e.g., forwarding, load balancing) and “application-level services” (caching, video transcoding, web server) seems largely driven by business model considerations. But it drives many architectural decisions which are not necessarily dictated from a technical perspective.

At the end of the day, this distinction has little justification on a technical level. Services comprise computational artefacts (containers, VMs, processes) and resources for running them (CPUs, GPUs, ...). It has little impact whether the computation is consulting a forwarding table or is transcoding a video file. Similarly, all services need communication, whether for actual data exchange or to update routing tables.

It seems that considerable added value could be achieved, in particular for operators, if they embraced the notion that both kinds of services could be orchestrated in the same fashion (and possibly even with the same orchestration infrastructure). The arguments in favour of (a) separating these types of services and (b) separating the orchestration infrastructure (one orchestrator vs. separate network/service orchestrators) appear to require continuous scrutiny as technology evolves. Architecture should be built which embraces this fact and is conducive to this approach.

5.3 Architecture options beyond the consensus architecture

The previous sections of this chapter have outlined what can be regarded as the common consensus on how to structure a MANO system. In particular, it lays the groundwork for the key interfaces and interactions of such a MANO system with its environment, opening the path to (already ongoing) standardization activities. Yet this consensus still has room for interpretation and differentiation among different realizations of such a system. This section explains some of the more prominent such open spaces, along with plausible variations for interpretation.

5.3.1 Structure of orchestration

5.3.1.1 Service-type: Integrated or segregated orchestration

For services, there is sometimes the distinction between “network-facing services” (i.e., services concerned with transporting packets in the narrower sense of the word, like packet forwarding, filtering, etc.) and more “application-facing services” (i.e., services that understand the semantics of the packets, like an add-insertion service or even just a webserver). While even this distinction is not uncontested, once we assume that this distinction makes sense and that such services can be reliably identified (or are explicitly tagged) as such, it raises a question for the orchestrator structure: Is there a single orchestrator that can deal with both (or even multiple) types of services (here called an “integrated” orchestrator), or does it make sense to have separate, specialized orchestrators that focus on only some of these services (here called a “segregated” orchestrator)?

There are obvious pros and cons to both approaches. Segregated orchestrators certainly make the architecture more complicated, introduce the need to assign areas of responsibilities from a resource perspective (which orchestrator is allowed control over which resources), solve the question of how to identify services pertaining to which orchestrator, possibly how to split a heterogeneous service description into its constituting “network” and “application-facing” parts,

and how to align control decisions taken by these two orchestrators (as there is a considerable danger of independent control algorithms not resulting in a desirable outcome). In an integrated orchestration approach, all these problems disappear.

On the other hand, an integrated orchestrator might turn out to be very complex if there is indeed a need to treat such different services substantially differently (a one-size-fits-all orchestration approach is indeed unlikely); the clear separation of areas of responsibility over resources might in fact be seen as an advantage for operational stability (e.g., a segregated RAN orchestrator could still maintain basic RAN services like phone calls even if an application-oriented orchestrator were to fail). There is a good argument that an integrated orchestrator is a more challenging piece of software (from both dependability and performance perspectives) but would result in a simpler architecture overall.

Examples for both approaches have been pursued in multiple projects. A solid comparison and a final verdict are still outstanding, though. In fact, from the perspective of the meta architecture, there is no need to standardize this option as both could be realized inside the meta architecture.

5.3.1.2 Flat vs. hierarchical orchestration

The previous question – integrated vs. segregated orchestration – dealt with service types and resource types and, consequently, whether there should be one or two (or even more) types of orchestrators. Orthogonal to that question is the question of whether there is only a single instance of a particular orchestrator type that is in charge of all assigned resources (a “flat” orchestrator) or whether there are multiple orchestrators (a “hierarchical” model, when orchestrators know they talk to each other). This is almost entirely a performance and scalability issue, to some degree a dependability issue as well. A hierarchical orchestrator is *not necessarily* a segregated orchestrator as all hierarchy members would deal with the same type of services.

A hierarchical model seems to be quite popular in current discussions. There are a couple of open questions in this context:

1. Is the number of hierarchy levels and the area of responsibility of each hierarchy member fixed up front (say, by a configuration action for a particular infrastructure)? Or can it be an auto-adaptive hierarchy, where upon load changes responsibility areas can be split/merged and new hierarchy levels can be added/removed and new orchestrator instances can be started/old ones stopped?
2. What is the *interface* between the orchestrators in such a hierarchy? In a flat model, an orchestrator’s NBI accepts service requests, and at its south bound, it talks to the NBI of an infrastructure abstraction (typically, a VIM). These two NBIs are structurally quite different. There are, hence, two options: Teach an orchestrator to talk to different NBIs (one being the NBI of a VIM, one the NBI of a lower-level orchestrator), necessitating to break up services into sub-services. Alternatively, an orchestrator could expose different NBIs, one being a conventional “service-style” interface, the other an infrastructure-oriented NBI. The advantage here would be that from the perspective of a higher-level orchestrator, it always talks to a VIM-style interface, making the idea of recursive orchestration much easier and elegant. In fact, an orchestrator does not really need to know whether it is talking to a true VIM or to another orchestrator pretending to be a VIM. This design choice has not yet been thoroughly investigated, to the best of our knowledge.
3. What is the relationship between siblings in such a hierarchy? Are they allowed to negotiate directly with each other (e.g., to “borrow” resources)? If so, this needs the definition of an east/west interface by which orchestrators on the same level can talk to each other, without being in a controller/controllee relationship.

Inside a single organization, such an east/west relationship complicates matters considerably without clear benefits; but in a cross-organization situation (sometimes

called “federation”), this could be the natural model to pursue instead of relying on a controlling, higher-level orchestrator in charge of multiple organizations (who would run that multi-domain orchestrator, why would that be an entity trusted to be impartial, what about competition among such multi-domain orchestrators, what would its business model be, etc.).

That appeal notwithstanding, mostly, the existence of one or several such multi-domain orchestrators is assumed, often in addition to such peering links between orchestrators on the same level. While the north-south-interfaces in this case are unchanged compared to the previous case, east/west interfaces are still under discussion.

4. Are hierarchy levels or areas of responsibility aligned with certain domains (in a very general sense of the word)? For example, is there a specialized orchestrator/control for an optical networking infrastructure, another one for computational resources, and another one for a wireless edge? This could be hierarchies along a technological domain. Another domain example are, naturally, organizational/company boundaries, in the conventional sense of “domain”. And last, a “domain” could be a subdivision of a larger infrastructure into an edge domain, a core domain, etc. (each one spanning multiple technologies, dealing with all kinds of services so not a segregated orchestrator in the sense of the previous section).

An example for such a domain-specific segregation is the use of two orchestrators (an NFV orchestrator and a MEC orchestrator) as demonstrated in [5-13]. A specific “dispatching layer” located on top of the two orchestrators receives all requests, performs the appropriate checks by looking into the descriptors and the details of the target slices, and forwards to the appropriate orchestrator if such a service deployment is possible. The component called “Multi-Tier Orchestrator” (MTO) is providing this functionality. This offers a simple interface for accepting “generic” service requests, which is an abstraction (or simple “forwarding”) of the NBIs (Northbound Interfaces) of the underlying orchestrators.

Again, trade-offs here are obvious but have not been thoroughly explored, nor have all the necessary interfaces been identified.

5.3.1.3 Orchestration vs. slicing

The relationship of an orchestration system and a slicing system is still not settled. This is not surprising as there is still, despite several years of frantic work, no commonly agreed upon definition of what a slice actually is; many definitions compete and they entail very different relationships to an orchestration system.

In a very straightforward view of slicing, a slice is simply a service with resource guarantees or guaranteed service level. This view fulfils most practical requirements towards a “slice”. In this view, the slicing system and the orchestration system are, of course, identical.

In a view that is more or less on the other end of the spectrum of opinions about slicing, a slice is a collection of resources – computing, networking, storage – that constitute a virtual network, embedded in some physical networking infrastructure. Inside such a slice, the slice owner has essentially full freedom to do what they like. For example, it could be a good idea to orchestrate services installed in such a slice and, to assist with that, an orchestrator could be installed inside such a slice. To assist with basic, fundamental services outside any slice (e.g., basic packet forwarding in the underlay), it makes sense to have yet another orchestrator installed outside of all slices. In this view, with N slices, there could be N+1 orchestrators running.

Yet another view effectively incorporates slicing functionality into the orchestrator (logically, implementation-wise it could be easily done as a separate subsystem) and triggers the creation of a new slice whenever a new service needs a slice (several services can share a slice). The slicing system has only the job of allocating guaranteed resources. Here, there would be a single orchestrator, in control of a single slicing system.

For all these approaches as well as many other combinations, some proponents exist and there is still no consensus in sight. Such a consensus is also unlikely to emerge until the definition of slicing is firmly settled. Also, many of the plausible options (e.g., a recursive approach where slices can be sliced again, each time with their own orchestrator) often earn criticism and adverse reactions that are ultimately rooted in a mistrust of a slicing approach to indeed firmly guarantee resource isolation and the unappeasable fear that resource consumption does spill over between slices. But this is difficult to circumvent on an orchestration level – either the slicing approach is trusted, then resources in a slice can be freely given to a tenant to do with as they please, or it is not trusted, then the entire approach seems superfluous and dangerous. Orchestration cannot quench that fear.

5.3.1.4 Abstractions and their violations

Irrespective of the way the relationship between hierarchy levels of orchestrators is organized, it makes a lot of sense to abstract and simplify the actually available resources on a lower layer when presenting capabilities to a higher layer. This will be necessary to obtain any scalability and performance benefits from hierarchies; it will also be necessary to work in a federated context when the peer orchestrator is not fully trusted and is not supposed to know the internal details of a domain and should rather obtain a condensed view only.

Such an abstracted view is a common idea. For example, in data centres, there is often the notion of a “big switch” abstraction, where the entire internal structure of the data centre’s network is ignored and, instead, a simple idea that all nodes are directly connected to a single, (very) big switch is used. For a data centre with abundant bandwidth and negligible delay differences between paths, this might be an acceptable simplification. But when trying to use an infrastructure to deploy NFV services for a highly distributed customer base, such an overly abstracted representation of a network seems counterproductive.

But it is also not clear what a good way to produce a simplified view of a network actually is. Let us consider a very simple example (Figure 5-2) of an actual infrastructure with four nodes, connected as shown in the figure (at the bottom) with each link having a data rate of 2 units. We want to present a simplified view of this setup to another orchestrator, only comprising nodes A and D, with a single link between them. Which data rate should we pretend that this virtual link can support? Option one (middle of the figure, left side) could be to use the sum of all data rates in the infrastructure network (more correctly put, the maximum flow from A to D), claiming a data rate of 4. Option two could be to just claim the minimal supportable data rate over a single path (which is actually not trivial to compute in the general case).

Suppose we now want to deploy a service consisting of three functions X, Y, Z onto this network, where X should be at A and Z at D. Suppose further that there are no resources left at A or D to also run Y; this service requires a data rate of 3 units from X to Y and Y to Z. Apparently, this could be mapped to the abstracted network using the first version – but this would violate requirements if the service’s network traffic is not splittable (or, equivalently, if function Y cannot be run in multiple instances due to statefulness). On the other hand, the second abstraction would refuse the service request, but that would be wrong if the service is indeed splittable (or function Y could be run in multiple instances owing to it being stateless). So, either of these simple approaches to abstract details of a network could be wrong, depending on the required service’s properties (here, splittable vs. non-splittable flows).

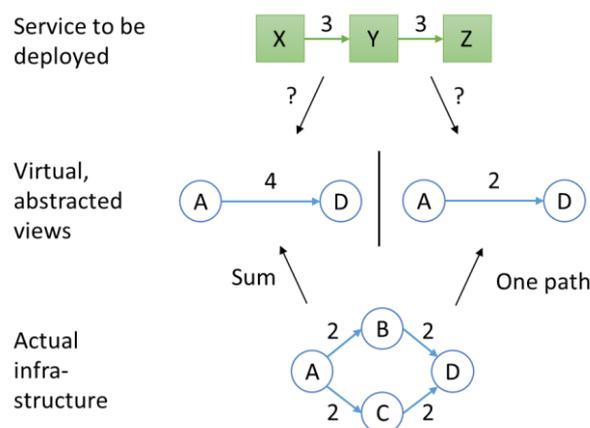


Figure 5-2 Different resource abstractions needed

Even more sophisticated approaches fail. For example, presenting a multi-graph with multiple edges between nodes as the abstracted model is tempting. But how then to deal with an infrastructure that has an additional edge from B to C? That means, there are three paths from A to D, but their capacities are shared. So, there is no obvious, simple answer here. In fact, this problem appears not to have received sufficient attention and requires further research.

5.3.1.5 Conflict resolution

With the increasing complexity of a 5G system, it will become increasingly difficult, if not impossible, to avoid conflicts between participating entities. Different types of conflicts can exist, for example:

- **Resource conflicts:** Several services have been accepted but they need the same resources to fulfil their quality promises. These conflicts can be due to incorrect admission control or overly aggressive oversubscription. While such a conflict will typically result in fines being paid, they still need to be resolved in a practical system.
- **Rules might conflict with each other,** for example, when composing a service out of functions that specify mutually incompatible packet forwarding behaviour. This can happen both in an NFV context or in an SDN context (imagine two different SDN applications running at an SDN controller's NBI, both being presented with the same PACKET_IN message and answering with mutually incompatible decisions, like "forward" and "drop"). Such conflicts can potentially be detected by a MANO system.
- **More generally, there can be *feature interaction* conflicts.** The previous rule conflicts can be seen as a low-level, packet-level example. At a service level, such a service conflict could be as simple as the classic "Call waiting" & "Call forwarding" feature interaction problem. As this is tied to application semantics (and in an NFV context, likely decided by some VNF), it is harder.

In all these cases, the conflicts need to be avoided (which is notoriously hard) or detected and resolved. In some cases, pre-fixed policies, either specified by the platform in general or by a service in particular, can help. Past experience, however, has shown that there are limits to this approach. A current research effort points towards learning such conflict resolution actions from inside an operational network, but there are no final results available yet.

5.3.1.6 Time scales

Most of the discussions so far have ignored timing aspects. Clearly, this cannot be ignored. We point out two aspects.

5.3.1.6.1 Short vs. long-term for slowly varying trends

It can make sense to separate short-term actions with very short required reaction times (e.g., actions on a flow level) from long-term planning actions (e.g., deciding where to run which function or how to scale a service). This can improve operational stability and performance. Such a separation can also be reflected by the architecture of a MANO system by splitting the MANO system into separate subsystems, each one responsible for different types of actions.

A typical terminology for such a split would be between “control” for short-time scale operations vs. “orchestration” for operations on longer time scales. This separation can be, but does not have to be mapped to the separation between a flow-level control entity (e.g., an SDN controller) and a service-/function-level orchestrator. The orchestrator then only has to deal with long-term trends, e.g., modifying the number of function instances during the course of a day.

While this separation is appealing from a software development and maintenance perspective (an SDN controller is a complicated enough piece of software), it does introduce yet another interface and operational dependency into an already complex architecture model. It also necessitates a decision on where to split overall functionality and decide which actions are short-term and which are long-term.

5.3.1.6.2 Dealing with load spikes

Even with short-/long-term separation in place, it could happen that traffic spikes occur which cannot be simply dealt with by operations in the purview of the short-term control system (i.e., if there are not enough function instances running, no amount of SDN rerouting is going to help). Hence, even the long-term orchestrator needs to be able to deal with short-term changes (which, in fact, calls into question the control/orchestration separation).

An important asset in dealing with spikes, irrespective of the MANO system’s architecture, is the ability to quickly bring up additional instances. The cloud computing answer to this problem is FaaS (or serverless computing), which allows to do just that, at low overhead: bring up functions on an as-needed, load-adaptive basis. However, this requires that the realized code is indeed a function, hence, stateless – there is no state maintained inside a function and it is not possible to move state between function instances. As long as this requirement is met, FaaS is indeed a promising option; however, it seems unlikely that all functions will be stateless. Applying FaaS does require that the MANO system understands the semantics of the functions constituting a particular service and understands which types of lifecycle management operations it can take.

5.3.1.7 Technologies

In principle, orchestration should be shielded from idiosyncrasies of underlying communication technologies, as is the basic tenet of a layered architecture. In detail, however, crossing layer boundaries might have advantages. We consider one example here when orchestration happens in the context of an optical/WDM network.

The NFV MANO architecture deployed for the advanced SDM/WDM fronthaul network integrates an NFV service platform for the management of network services and network slices for verticals, a transport SDN controller operating the optical fronthaul network, and an edge computing controller allocating computing and storage resources in the central office (CO). At a high level of abstraction, this architecture uses typical roles and separation of concerns (NFVO, VNFM, PNFs, VIMs; the NFVO is also in charge of slices). The NFVO coordinates the configuration of both VNFs and PNFs through the VNFM and PNFM. The control of the optical fronthaul network is delegated to the transport SDN controller, but still under the global coordination of the NFVO. The communication between the NFVO and the transport SDN controller is based on the transport API (TAPI) specification, with extensions to deal with the specific optical technologies of the fronthaul network. The transport SDN controller is extended

with a dedicated transponder manager to interact with the analogue radio over fibre (ARoF) transceivers and optical beam-forming network (OBFN) system (both at the CO and the radio unit (RU)) through dedicated agents using a REST API. The NFV service platform is extended with a network slice manager to deliver multi-tenant virtual environments. It is deployed as a client of the NFVO and manages the life-cycle of network slice instances. It translates the vertical requirements into a suitable NFV network service with the required QoS, which is instantiated and terminated at the NFVO. A client of the NSM is the mobility and power manager. On the one hand, it is responsible for the life cycle management of the optical power channels between the CO and the RU, and on the other, it implements user mobility logic, such as decisions about activation and deactivation of femtocells based on the users' location, coordinating this with the vertical service instantiation procedures.

5.3.2 Implementation patterns for MANO frameworks

A number of common implementation patterns for MANO systems has emerged. One is that of the *monolithic orchestrator*. In the reference architecture, an orchestrator has a lot of responsibilities. Realizing all those in a single, monolithic piece of software might be feasible, but seriously jeopardizes maintainability, dependability, and performance. Hence, more suitable implementation patterns are needed.

As discussed in Section 5.3.1.6 from a time-scale perspective, splitting an orchestrator into a controller for short-time actions and an orchestrator proper for longer-time actions is a promising first step. This *coarse-grained functional split* is pursued by multiple projects that do not specifically focus on orchestrator implementation patterns [5-14][5-15].

To improve flexibility and to ease implementation of such a complex piece of software, the software engineering community has developed multiple approaches. One of these approaches is based on the notion of microservices, connected by a software bus that realizes a publish/subscribe paradigm between its components. Applying this concept to an orchestrator leads to a much finer-grained functional split, where individual functional boxes can deal with separate aspects of a request pertaining to a function or a service (e.g., to turn off a service) [5-16]. A strength of this approach is how easy it is to extend it; for example, a data analysis framework was easily added to this orchestrator. It also did lend itself nicely to the incorporation of slicing support or support of multiple networking technologies.

Such a microservice-based orchestrator is not tied to a single machine. Provided a suitable, well-performing pub/sub system was chosen, a distribution of the orchestrator's components across multiple machines for improved dependability and performance is easily possible. Pub/sub or software buses are usually based on top of existing open-source projects (e.g., Kafka [5-26] or RabbitMQ [5-27]) with well-tested performance.

It is possible to take the flexibility and simple extensibility of such a service-based platform one step further. This feature has proven advantageous specifically in the VNF and service management domain, as management of VNFs and network services is highly specific, e.g. for configuration and scaling actions, which are highly dependent on the specific functions and the environment they are running in. To this end, the service developer can ship the service package to the service platform together with service- or function-specific lifecycle management requirements and preferences, called Service-Specific Managers (SSM) and Function-Specific Managers (FSM), respectively. SSMs and FSMs can influence the Service and VNF lifecycle management operations, e.g., by specifying desired placement or scaling behaviour [5-17]. By virtue of a modular design in the Management and Orchestration Framework of the service platform, the service platform operator can customize it, e.g., by replacing the conflict resolution or information management modules. This idea is illustrated in Figure 5-3.

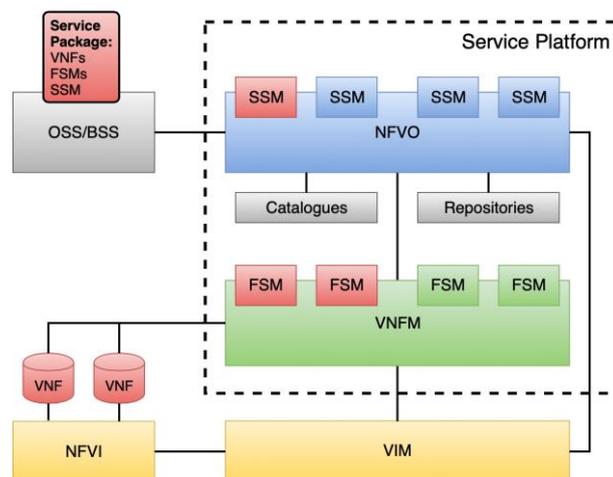


Figure 5-3 Service-Specific Plug-In Extensions

Given the popularity of such message-bus based orchestrator structures in several projects, it seems like the go-to solution today. There are, however, some possible disadvantages that should be taken into consideration before deciding for that approach. One is the difficulty to provide real-time operation (message busses are, by definition, decoupled in time); the second are scalability concerns (albeit systems like Kafka or RabbitMQ are known to easily scale to millions of events per minute, which should be enough for service lifecycle management actions even in a large-scale orchestrated network). Another potential downside is that debugging becomes more difficult. One reason for this is the distribution of components and the associated effort in gathering information across a potentially large number of components at multiple locations. More important, though, is that interactions between components are not directly obvious. Looking at these interactions requires additional tools which track, correlate and visualize communication between components via message buses and, potentially, other technologies, such as REST-based interactions.

5.3.3 Algorithmic building blocks

Every orchestrator/MANO system has a multitude of algorithmic decisions to take. Which algorithmic problems have to be solved is mandated by the orchestrator's structure and tasks; how they are realized, on the other hand, can vary widely and could be an opportunity for vendor differentiation.

5.3.3.1 Example building blocks

To name but a few typical algorithmic questions that need to be solved:

- **Placement:** Which instance of a service's functions should run on which resource, supported by how many resources (like CPU clock). Note that resources here might be both actual or virtual resources; in the case of virtual resources, this turns into a recursive problem. Placement is closely related to the NP-hard facility location problem.
- **Routing between functions of a chain:** Once function instances are placed, routes need to be found between them to connect them in the right order. Algorithmically, this is a multi-commodity flow problem, which is also NP-hard. Together with placement, this is often considered as the virtual network embedding problem, where the virtual network to be embedded is the service's application graph.
- **Scaling:** Services become interesting when they can adapt to load changes by spinning up new function instances. Hence, we cannot just embed a fixed virtual network or a fixed service graph; rather, we need to dynamically scale these graphs before embedding them

to ensure the right number of instances is available to support a given load. Of course, this is also NP-hard.

- **Load/traffic prediction:** How to predict traffic and load changes, to give an orchestrator sufficient time to react? Similarly, how to predict failures? (Of course, predict in a stochastic sense.)
- **State management between VNFs:** Assuming that VNFs are stateful and they should be scaled up or down, what is the best combination of state management actions to lead to the smallest overhead?
- **Monitoring:** Which data should be monitored, where and how often to obtain the best possible approximation of ground truth at the smallest possible overhead?

5.3.3.2 Realizations, MAPE-K example

Many of these building blocks are some form of an optimization problem, more or less well hidden. For some of them, a conventional formulation as some variant of an integer problem (mixed, quadratic, etc.) is often possible and convenient, directly using a solver (from simple open-source solvers like GLPK [5-28] to commercial solvers) in a deployment context. Solving times are often too long and, ideally, an approximation algorithm can be found (often with orders of magnitude better runtimes yet still guaranteed performance ratios). In absence of such approximation algorithms, one typically has to take recourse to mere heuristics (again much faster than an optimizer but, unlike an approximation algorithm, without performance guarantees). The advantage of heuristics is that they usually can be developed when other attempts fail, and they often can come in both centralized and distributed versions.

One currently (again) popular way to structure such heuristics is to take up the idea from autonomic computing. There, the idea was to structure such problems into several phases, namely Monitoring (observe the actual state of the system to some feasible and desired degree), Analysis (derive more compact representations of that state; representations about which can be argued), Planning (deriving desirable state changes and deciding which actions can push the current system state towards such a desirable state) and finally, Execution of such planned actions in the actual system. These four phases are often executed in a continuous loop, hence the name MAPE loop. Incorporating an additional knowledge base where, e.g., prior state observations, the taken actions and the resulting state changes are stored, leads to MAPE-K. Knowledge bases are typically updated during operation in order to improve decisions.

In the simplest case, this MAPE or MAPE-K loop is indeed confined to a single algorithmic black box. If actions to be executed have impact on the status of other boxes or execution needs to be better coordinated, MAPE may span across boxes and becomes a basic design decision of an orchestration framework.

Depending on the concrete implementation structure of the orchestrator, integrating such a MAPE-K approach can be quite easy. In particular, microservice/bus-based approaches as outlined above nicely lend themselves to such an extension, where MAPE-K algorithmic boxes can be easily integrated. Depending on the scope of the individual decisions and the scope of knowledge, an entire MAPE-K loop can be encapsulated into a single box connected to a message bus; it is also easily conceivable to factor out, e.g., the knowledge component, connect this separately to the message bus, and use this to interact with multiple MAPE boxes [5-18][5-19] (the event-driven nature of message buses lends themselves very nicely to that approach as the knowledge component can generate events to the other, subscribed components). Execution components can then either use the message bus again to send commands to actual low-level interfaces or take actions themselves.

A possible downside of this approach (and partially inherited from the message bus approach) is that real-time operation is difficult if not impossible to ensure. It needs very careful design to ensure that all components of such a MAPE-K approach indeed perform all their operations in

bounded time, irrespective of load levels. Often, however, “real time” (in its strict interpretation of “absolute guarantee of all execution times, even in the presence of failures”) is an overly ambitious requirement or goal and “near real time”, “soft real time” or “rather fast” are good enough.

An interesting option exists to structure such MAPE-K systems in a hierarchical manner. This is mostly an algorithmic design issue as long as the hierarchy of MAPE-Ks is hosted inside a single, message-bus based orchestrator – all messages can easily flow between them in the conventional fashion. It gets more interesting if MAPE-K boxes are distributed across orchestrators that are themselves hierarchically structured. Then, it is likely not advisable to spread the message bus across multiple orchestration instances (even if technically perhaps feasible). Likely, a separate protocol between these MAPE-K instances will have to be defined; this is not clearly established yet and would likely need a standardization effort to be useful.

5.3.4 Description mappings

In a MANO framework, descriptions exist for many types of artefacts: from infrastructure, to functions, to services, to slices, policies, SLAs, tests, and possibly to business objectives. There is often a need to map between different description formats, both horizontally, i.e., between description formats for the same artefacts, and vertically, meaning from abstract to concrete.

5.3.4.1 Horizontal mapping

The descriptions in a MANO framework can be quite heterogeneous, e.g. because underlying systems (like a VIM) need different description formats. It becomes desirable to support multiple formats of such descriptions. A viable approach is to translate descriptions from one format to another, instead of creating the $n+1^{\text{st}}$ “standard” description formalism.

ETSI recently started to define and specify a common VNF package format, based on the TOSCA CSAR standard. Even though this package format is a good starting point, some important features are still missing in the specification, e.g., support for complete network services inside a package. Ideally, a generic package format emerges which allows packaging VNFs and services for different target platforms to simplify on-boarding procedures on different infrastructures as much as possible.

Some extensions to descriptors which are currently being explored include:

1. *Recursive NSDs* that allow the description of recursive network services, i.e., including not only VNFs but also other network services. Allowing such recursive references enables faster and easier reuse and extension of existing network services by just reusing and referencing the corresponding NSDs. In doing so, the creation of new NSDs also becomes less error-prone.
2. *Layered descriptors* which allow to package different descriptor formats for the same VNF or network service inside a single package, e.g., SONATA and OSM descriptors describing the same VNF or network service. The benefit of this concept is that a developer can ship a service that is compatible with different platforms within a single package. Tool support can be developed which automatically creates multiple formats from a common code or descriptor base.
3. *Test result packages* which are created and signed by a validation and verification provider. Besides the actual test results, these packages reference the exact service or VNF for which the tests were executed. The signed test result packages attest certain service properties, e.g. specification compliance or QoS levels. The signature allows verification of the integrity of the package and its contained attestations.

5.3.4.2 Vertical mapping

Vertical service blueprints and vertical service descriptors (VSD) can be used to describe vertical services including their SLA requirements [5-14]. The SLA requirements can be of different kinds, for example:

- end-to-end latency and bandwidth requirements, necessary for the service to function correctly,
- number of supported users, coverage area, etc., related to the dimensioning of the service,
- availability and reliability,
- deployment time, energy efficiency, i.e., optimization targets for the deployment of the service.

The vertical service description including such SLA requirements can be translated to an NSD with appropriate selection of deployment flavours and instantiation level. To that end, the NSD is determined from the VSD. Using a rule-based approach, specific values of the SLA requirements are then translated into the selection of deployment flavour and instantiation level. Some of the SLA requirements can be encoded in the NSD itself. Continuing the example above:

- bandwidth requirements can be expressed in the NSD as bandwidth requirements on virtual links,
- the number of supported users can be mapped to a corresponding instantiation level with sufficient VNF instances handling the expected number of users,
- reliability can be mapped to a deployment flavour with or without redundant components,
- energy efficiency can be mapped to an orchestration policy for the NFVO to place VNFs in the most energy efficient way, trigger activation of corresponding features, even if these imply license fees.

Additionally, a vertical or other customer may request instantiation of several services and they may have agreed on an overall resource budget regarding compute, storage, transport, and radio capacity with the provider. In case the resource budget is insufficient for all instances, there has to be an arbitration among the services. This arbitration should take the service priority into account and might imply a change of the deployment flavour and/or instantiation level of services. Once these steps are performed on the NSDs, they can be instantiated.

5.3.5 Monitoring aspects in Orchestration

The monitoring system in a MANO framework needs to monitor all virtualized resources and, through an appropriate set of parameters, the applications and services running on the infrastructure. Regarding the infrastructure, the monitoring system typically includes three different resource domains, including 1) NFVI resources; 2) SDN-enabled elements; 3) physical devices that do not belong to the first two categories. With regard to the applications and services, the monitoring system includes VNFs and service monitoring parameters and metrics, useful also to check SLA compliance.

The monitoring system is integrated with the different orchestration layer components to assist in network and systems management and to provide a coherent and simple-to-access view of the platform exposed to both dashboards and analytical techniques. It collects all the information to enable a “monitoring as a service” model.

The set of the monitored parameters can include VM-related information, e.g. CPU utilisation, bandwidth consumption, as well as VNF specifics such as calls per second, number of subscribers, number of rules, flows per second, VNF downtime, video streaming start success ratio, video streaming start delay, video streaming stall frequency, video streaming download throughput, etc. One or more of these parameters, depending on the implemented logic, could also trigger a

reaction in the QoS loop. At the service level, monitoring parameters represent metrics that are tracked to check the level of compliance with the active SLAs.

In many projects, Prometheus [5-29] has been selected as open source package to fulfil the monitoring system requirements, along with Grafana [5-30] for data analytics and visualization. As not everything can be instrumented directly, applications that do not support Prometheus metrics natively can be instrumented by using exporters. The use of exporters allows collecting statistics and metrics converting them to Prometheus metrics.

5.4 DevOps meets Orchestration

NFV as such can already be seen as an embodiment of the microservices approach. With the discussion above, not only the services themselves, but also the orchestration components can be considered as microservices. This approach brings big advantages, including flexibility, continuous delivery (CD) and integration, reduced time-to-market (TTM) and time to deployment, faster resolution of problems, more stable operating environments, improved communication and collaboration, reduced costs and higher dependability, etc., but requires a new approach to development and operation.

A common theme in the cloud and software industries is DevOps – the integration of development and operation of complex software systems. Clearly, this is a strong candidate for NFV and orchestration as well, and it is pursued under different perspectives [5-20][5-21]. It is commonly acknowledged that an efficient DevOps approach crucially depends on appropriate support tools.

Specifically, tool support is needed at multiple stages (for both services and orchestration software), which interlink with each other: development time, pre-deployment time, deployment time, and runtime. So far, this chapter has mostly focused on runtime support functions and tools. This section focusses on the phases prior to runtime.

Depending on the concrete orchestration approach, the DevOps approach also changes the internal structure of an orchestrator. An example architecture that considers these needs is proposed in [5-21], which distinguishes three different frameworks: CBTR (from Coding-Building-Testing-Releasing), Monitoring, and Management.

5.4.1 Development time: SDKs

At development time, software development kits (SDKs) are needed. In the NFV world, they should complement general-purpose SDKs for generic programming tasks by supporting NFV-specific needs. They should comprise support for service requirements, service design, and specific implementation tasks [5-13][5-22]. Usually, these SDKs are provided as stand-alone tools (rather than being integrated into, say, Eclipse, which would reduce their practical appeal) and available as a collection of command-line tools for universal use, sometimes endowed with a GUI. In some cases, extensions to editors (by providing specifications for a template as a domain-specific language) or standalone editors are made available as well. For example, support for template development exists, both for functions and services. Different semantics are supported, from plain ETSI semantics to more expressive ones (e.g., allowing to specify traffic-dependent resource consumptions or monitoring points).

Some attempts are made to support non-networking developers (e.g., members of a “vertical” tenant) by simplified descriptions and models and, possibly, by specific tools like a service composer (which produces similar artefacts as an editor, yet in a simplified usage environment). This is, however, a double-edged sword as it by definition limits universality, and costs for such tools need to be amortized over a much smaller number of cases.

5.4.2 Need for validation tools

The artefacts that exist in an orchestrator are numerous – descriptions of functions, services, infrastructure, SLAs, etc. It is inconceivable that all these artefacts, if produced manually, were free of errors. Hence, even on a simple, syntactic level, there is a need for specific validation tools. These tools check, e.g., consistency of XML files, completeness of descriptions, etc. On a semantic level, they can check whether all required artefacts (e.g., function implementations) are available, whether all required connections between functions are indeed stated, or whether some function's gate has been left unconnected [5-22].

Such tools are commonly used in many projects. Moreover, there is a class of tools that can validate the correctness of both functional and non-functional claims about a function or a service (within plausible limits; in its full generality, this would resort to solving the incomputable halting problem).

All these validation tools can be used at different points in time. They are useful during evaluation time and deployment time when on-boarding a service. They are also useful, in addition, at a pre-deployment phase when a tenant intends to deploy a service on a (virtual or actual) infrastructure and the infrastructure provider needs to check claims about required resources to achieve a desired service quality level at a given traffic level [5-20]. Such tools are crucial for an infrastructure provider in order to ensure service quality (and to be able to sign contracts on such assurances).

It is important to realize that the need for such tooling is aggravated in a 5G environment compared to a mere cloud environment. This is partially due to the complex relationship of a network operator to possibly many software vendors introducing a high diversity into a system for which stability is imperative.

5.4.3 Need for evaluation tools

Practically speaking, a validation tool on its own only checks syntactic and semantic properties. Checking quantitative and performance claims is difficult in a general sense. It makes sense to outsource this to a separate family of tools, here called evaluation tools.

These tools have the task to either derive or validate performance claims. This is not a simple task in general; performance prediction of arbitrary code is a difficult problem in software engineering in general. A plausible approach here is to use tools that provide a desired environment and subject the service/function of interest to different load levels, e.g. to different levels of traffic. Such an environment can be actual (very costly and maintenance-intensive, but quite accurate and suitable as a staging environment for actual operation) or virtual, using emulation techniques with reasonable accuracy at much smaller resource demands [5-31].

It should be pointed out that these evaluation tools can be used at development, pre-deployment and deployment time, with different foci and likely by actors in different roles. At pre-deployment time, for example, a neutral "Validator" role can take a service package along with a performance claim stated by its vendor and validate it using its own evaluation tools, providing an authenticated witness to such claims (or a counterexample refuting a claim).

It should be noted that such an approach, along with several levels of staging areas, is quite common in the cloud computing community and is starting to be recognized by the networking community as useful and beneficial. Various concrete workflows for how to integrate such tooling have been proposed and are not yet aligned.

As a concrete use case beyond validation, such evaluation tools can also be used to prime a knowledge pool in a MAPE-K context with observations about a function's or a service's behaviour. This can help getting a MAPE loop started without having to rely on an empty knowledge pool, but still gives a natural venue of updating such knowledge with actual observations made during deployment time of a service.

5.5 References

- [5-1] ETSI. Network Functions Virtualization (NFV). URL: <https://www.etsi.org/technologies/nfv>.
- [5-2] ETSI. Zero touch network & Service Management (ZSM). URL: <https://www.etsi.org/technologies/zero-touch-network-service-management>.
- [5-3] ETSI. Experiential Networked Intelligence (ENI). URL: <https://www.etsi.org/technologies/experiential-networked-intelligence>.
- [5-4] ETSI. Multi-access Edge Computing (MEC). URL: <https://www.etsi.org/technologies/multi-access-edge-computing>.
- [5-5] ETSI. ETSI GR MEC 017: Mobile Edge Computing (MEC); Deployment of Mobile Edge Computing in an NFV environment. Available online: https://www.etsi.org/deliver/etsi_gr/MEC/001_099/017/01.01.01_60/gr_MEC017v010101p.pdf.
- [5-6] ETSI. ETSI GS NFV-INF 003: Network Functions Virtualisation (NFV); Infrastructure; Compute Domain. Available online: https://www.etsi.org/deliver/etsi_gs/NFV-INF/001_099/003/01.01.01_60/gs_NFV-INF003v010101p.pdf.
- [5-7] ETSI. ETSI GS NFV-INF 004: Network Functions Virtualisation (NFV); Infrastructure; Hypervisor Domain. Available online: https://www.etsi.org/deliver/etsi_gs/NFV-INF/001_099/004/01.01.01_60/gs_nfv-inf004v010101p.pdf.
- [5-8] ETSI. ETSI GS NFV-INF 005: Network Functions Virtualisation (NFV); Infrastructure; Network Domain. Available online: https://www.etsi.org/deliver/etsi_gs/NFV-INF/001_099/005/01.01.01_60/gs_NFV-INF005v010101p.pdf.
- [5-9] ETSI. ETSI GR NFV-EVE 012: Network Functions Virtualisation (NFV) Release 3; Evolution and Ecosystem; Report on Network Slicing Support with ETSI NFV Architecture Framework. Available online: https://www.etsi.org/deliver/etsi_gr/NFV-EVE/001_099/012/03.01.01_60/gr_NFV-EVE012v030101p.pdf.
- [5-10] 3GPP TS 23.501, “System Architecture for the 5G System”. Available online: <https://www.3gpp.org/DynaReport/23501.htm>.
- [5-11] 3GPP TS 28.801, “Telecommunication management; Study on management and orchestration of network slicing for next generation network”. Available online: <https://www.3gpp.org/DynaReport/28801.htm>.
- [5-12] 5G Essence Project. Deliverable D2.1: System Use Cases and Requirements. November 2017.
- [5-13] 5GCity Project. Deliverable D4.1: Orchestrator design, service programming and machine learning models. June 2017.
- [5-14] 5G-TRANSFORMER Project. Deliverable D1.3: 5G-TRANSFORMER refined architecture.
- [5-15] 5G-PICTURE Project. Deliverable D2.2: System architecture and preliminary evaluations. July 2018.
- [5-16] 5G-MoNArch Project. Deliverable D2.3: Final overall architecture. April 2019.
- [5-17] SONATA Project. Deliverable D2.2: Architecture Design. December 2015.
- [5-18] 5G-MEDIA Project. Deliverable D3.3: Specification of the 5G-MEDIA QoS Control and Management Tools. August 2018.
- [5-19] NRG5 Project. Deliverable D1.2: NRG-5 Reference Architecture & Functional Decomposition. May 2018.
- [5-20] 5GTANGO Project. Deliverable D3.3: Final release of the VNFs/NS Store. May 2019.
- [5-21] NGPaaS Project. Deliverable D2.2: Final overall Next Generation PaaS Vision. January 2019.
- [5-22] 5GTANGO Project. Deliverable D4.2: Final release of the service validation SDK toolset. May 2019.

- [5-23] Linux Foundation Projects. ONAP – Open Network Automation Platform. URL: <https://www.onap.org>.
- [5-24] ETSI. OSM – Open Source MANO. URL: <https://osm.etsi.org>.
- [5-25] 5GTANGO Project. SONATA powered by 5GTANGO. URL: <https://sonata-nfv.github.io>.
- [5-26] The Apache Software Foundation. Apache Kafka. URL: <https://kafka.apache.org/>.
- [5-27] Pivotal Software, Inc. RabbitMQ. URL: <https://www.rabbitmq.com>.
- [5-28] GNU project. GNU Linear Programming Kit. URL: <https://www.gnu.org/software/glpk/>.
- [5-29] Cloud Native Computing Foundation. Prometheus. URL: <https://prometheus.io>.
- [5-30] Grafana Labs. Grafana. URL: <https://grafana.com>.
- [5-31] Open Source MANO. VIM emulator. URL: https://osm.etsi.org/wikipub/index.php/VIM_emulator.

6 Deployment, Evaluations & Analyses

6.1 Introduction

This section provides an analysis of how the projects in 5G PPP phase II deployed the overall architecture in the context of the use cases of the vertical use cases that each project addresses. It is an attempt to evaluate how the projects used certain concepts that are central to the 5G PPP system, such as virtualisation, functional split at the Radio Access segment or Multi-access Edge Computing capabilities. The evaluation should help to understand how the performance Key Performance Indicators (KPIs) have been considered and measured.

6.1.1 Performance KPIs

The technical Annex to the 5G PPP contractual arrangement [6-1] defines the following KPIs:

- Providing 1000 times higher wireless area capacity and more varied service capabilities compared to 2010.
- Saving up to 90% of energy per service provided.
- Reducing the average service creation time cycle from 90 hours to 90 minutes.
- Creating a secure, reliable and dependable Internet with a “zero perceived” downtime for services provision.
- Facilitating very dense deployments of wireless communication links to connect over 7 trillion wireless devices serving over 7 billion people.

These KPIs have been refined in the course of the execution of the 5G PPP programme in various white papers, among others in “5G empowering vertical industries” [6-2]. A more detailed and partly formal definition of the KPIs that are relevant for the performance of the 5G system have been defined by standards bodies such as ITU-T and 3GPP.

Report ITU-R M.2410-0 (11/2017) defines KPIs specific to the radio interface. These include *Peak data rate*, *User experienced data rate*, *Mobility*, *Latency* – separately for user plane and control plane, *Connection density*, *Reliability*, *Area traffic capacity*, *Peak spectral efficiency*, *5th percentile user spectral efficiency*, *Average spectral efficiency*, *Energy efficiency*, *Mobility interruption time* and *Bandwidth*.

In TS 28.554 [6-3], 3GPP specifies end-to-end Key Performance Indicators (KPIs) for the 5G network and network slicing. 3GPP introduces KPI categories; *Accessibility*, *Integrity*, *Utilization*, *Retainability* and for future updates also *Availability* and *Mobility*. The categories are defined with reference to ITU-T Rec.E.800 [6-4].

Accessibility refers to Registered Subscribers of Network and Network Slice Instance through AMF and UDM, Registration success rate of one single network slice instance, as well as Data Radio Bearer (DRB) Accessibility for UE services. *Integrity* refers to End-to-end Latency of the 5G Network, Upstream/Downstream Throughput for network and network slice instance, Upstream/Downstream throughput at N3 Interface (between RAN and UPF) as well as throughput between RAN and UE. *Utilization* refers to the Mean number of PDU sessions of network and network Slice Instance and the Virtualised Resource Utilization of Network Slice Instance. Finally *Retainability* refers to QoS flow Retainability.

Furthermore NGMN published a Testing Framework for the NGMN 5G pre-commercial network trials. Among others this paper specifies general requirements for testing, deployment scenarios, trial setup requirements, trial test requirements and service or technology specific requirements for several identified KPIs, such as *Latency*, *User throughput*, *Cell Capacity*, *Spectral Efficiency*, *Coverage*, *Mobility*, *Reliability and Retainability*, *User Experience*, *Energy Efficiency*, *Inter-RAT*

procedures, RAN architecture split, as well as *Location/Positioning service* and *Fixed Wireless Access*.

As can be derived from the main references above, there exists a large number of KPIs with partly diverging definitions, although these definitions are being consolidated by the standards bodies and the industry. The ad hoc work group of the 5G PPP has made an attempt to provide a consolidated view of the KPIs that are being addressed by the various projects of the 5G PPP programme.

6.1.2 Types of evaluation and usability of KPIs

Performance indicators are generated to monitor the quality of services, applications or networks that offer services to users to which we must guarantee enough quality so that they can offer or deploy their services. However, it is necessary to classify the type of deployment in which we want to generate these indicators depending on the state of deployment. We can establish a first high level classification in three levels:

- **Phase 1: Interoperability** of elements and adaptation to standards. In this phase, the main objective is to verify that the different devices interact in the expected way and conform to established standards. Normally in this phase the UEs are validated and the elements are configured to eliminate any interoperability problems. In this phase, the use of instrumental equipment is usually required to perform low-level interoperability reports and exhaustive compatibility validations of different measures.
- **Phase 2: Proof of concept and scalability** before a real deployment. This phase is prior to the deployment of networks and services, and the main objective is to prove that the service and objective applications of the service can be functionally provided. We also want to check to what extent the system is scalable and can support load tests. Equipment with load simulators or automatic equipment to perform functional tests are the ones that are most commonly used in this phase.
- **Phase 3: Monitoring in service** once deployed. This phase allows monitoring the quality of the services once they have been deployed in the real network, and they require periodic reports that allow taking preventive measures and controlling the quality that is really being delivered to the end users. In this phase, many data are integrated in the most efficient way possible, and usually require the deployment of components in the applications that generate part of the information used to monitor the quality of service in the users. This phase generates a lot of information, and it is critical to provide efficient data aggregation mechanisms.

When we talk about KPIs it is good to have as reference what is the objective in which we are considering their generation, since that allows us to use in the most efficient way instrumental equipment, load simulation equipment or possible applications development.

In Phase 1, interoperability tests of the UEs with the networks will be carried out, depending on the features and frequencies of the network that are being deployed, it is necessary to repeat these tests to ensure that we are within the parameters and configured expected and defined in the standards

In Phase 2, load and simulation tests will be performed in laboratory environments that will allow us, before making a real and massive deployment of the service for vertical companies, to optimize the resources used efficiently. In this phase we will verify that we can reach our goal of coverage and concurrency that have been established.

Phase 3 is the most critical and complicated, since having well controlled during the operation of the quality of service provided to users will allow us to anticipate any problem and in this way we can guarantee a better quality of service.

Any Vertical Industry will have to go through these three phases, and therefore the KPIs that will be implemented, tested and measured, should be applied in the most efficient way to each of these three phases.

6.1.3 Approach/options to generation of KPIs

In the context of KPI validation by the 5G PPP project an abstract 5G system partitioning has been used as illustrated in the figure below. The figure intentionally simplifies certain details of the architecture that pertain to the concept inside the presented segments, such as the functional split at the radio segment.

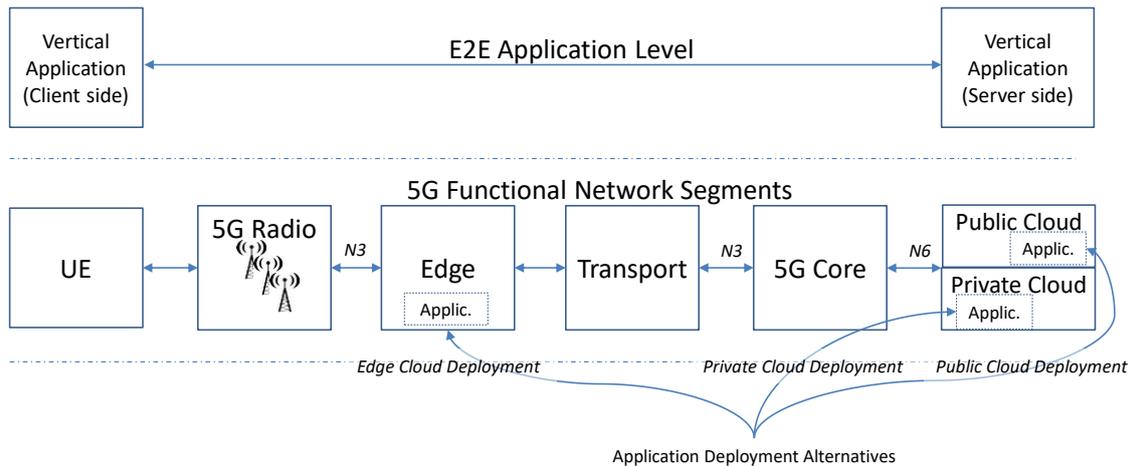


Figure 6-1 Functional Network Segments of a 5G System

For the analysis of the deployment options chosen for the architecture validation the following parameters have been considered:

- MEC deployment location. This parameter considers the architectural option to use computing capacity near the application for meeting performance requirements such as latency
- Functional split. This parameter considers the architectural option to implement functional split at the RAN for optimising performance parameters such as throughput and latency.
- Cloud RAN. This parameter considers the architectural option to implement different Cloud RAN options for optimising performance parameters such as throughput and latency, as well as cost, complexity, and energy consumption
- Transport technology. This parameter is used to assess the impact of the used transport technology mainly at the wide area network segment
- Spectrum used. This parameter is used to assess the impact of the use of different 5G spectrum option when available
- Implementation of SA/NSA. This parameter is used to assess the readiness of the technology to implement a full 5G system.
- EPC/5G core. This parameter is used to assess the readiness of the technology to implement a full 5G system and the impact of using 4G components.

Some of the results below are based on pre-standards prototype implementations of the different components that comprise a 5G system according to 3GPP.

6.1.4 Architecture instantiations

The analysis of the instantiations of the architecture has been performed based on the following questions that were answered by all projects.

Do you use a MEC deployment? Can you make a statement about its location?

This question interrogates whether the concept of multi-access edge computing has been deployed and used and potentially at which location. A significant number of projects have used MEC

Table 6-1 MEC deployment

	Use case	5G service class
Yes	In building premises, Virtual gateway, Port, Touristic city, Stadium (x2), Airplane, Emergency (x2), Smart city (x4), Predictive maintenance, Smart grid (x2), Smart museum, Automotive V2X (2x), Satellite edge (CDN), eHealth	eMBB, URLLC, mMTC
No	Media Distribution, High Resolution Media, Smart Optical Infrastructure, Smart City (Lighting), Dense Urban, Touristic city, Hotspot, Manufacturing, eHealth, Media	eMBB, mMTC
Close proximity, no explicit MEC	Smart Manufacturing, Immersive Media, Communications Suite	URLLC, eMBB
Not applicable	Railway, Distribute System Testing	(URLLC)

It is notable that the use cases related to media – such as immersive media, media distribution or high resolution media – did not deploy or use the MEC capabilities of the 5G architecture. In the case of 5G/satellite integration MEC capabilities have been deployed at the edge for pushing, pre-fetching and caching content. This conclusion is in line with the general assumption that the eMBB service class, not being sensitive to latency, will generally not use MEC. In contrast smart city use cases have used both options. This fact indicates that the notion of “smart city” is including a rather broad set of different applications which may or may not need a MEC deployment.

Do you use functional split? Can you make a statement about where precisely?

This question interrogates whether functional split has been implemented and deployed and where. Although this question implied the use of functional split at the radio segment, some the projects misinterpreted this question referring to the split of control and data planes, the function placement, or the separation of network services per target application.

Table 6-2 Functional split of radio segment

Functional split option	Use case
Option 2 F1, PDCP-RLC split (x4) most common CU/DU split	Smart city, Automotive, eHealth, Media, Manufacturing
Option 4 RLC-MAC split (x2)	Port, Touristic city,

Option 7 intra-PHY split (x2)	Smart city, Automotive, eHealth, Media, Manufacturing
Option 7 intra-PHY split with analogue radio over fibre	Dense urban, Hotspot, Touristic city
Option 8 PHY-RF split	Automotive, eHealth, Media, Manufacturing
Flexible degree of CU-DU split	Smart grid, eHealth, Smart city
No (x7)	Media, Automotive, Smart city (Lighting), In building premises
Not Applicable (x4)	Railway, Stadium, Smart Optical Infrastructure, Distributed System Testing,

Concerning the functional split at the radio segment a small number of projects have implemented and used the concept. It can be derived that the most commonly used functional split is option 2 according to the 3GPP functional split options recommendation as described in section 3.1.1. Further options that have been implemented and explored include option 4 RLC-MAC split, option 7 intra-PHY split, option 7 intra-PHY split with analogue radio over fibre and option 8 PHY-RF split. Implementation of, and experimentation with several flexible CU-DU split options has been conducted by at least two projects. There is no evident mapping of the implemented functional split options to use cases or 5G service classes.

Table 6-3 Other interpretations of functional split

Not related to radio segment	Use case
Network isolation	Manufacturing
Flexible function placement (x4)	Immersive media, Satellite edge (CDN), Emergency, High resolution media,
Control and data plane separation (x2)	Stadium, Airplane, Emergency, Predictive maintenance, Smart Grid, Smart city

Further interpretations of functional split pertain mainly to network isolation, flexible network function placement, as well as control and data plane separation.

Do you use Cloud RAN?

This question interrogates the use of RAN processing in a virtualised computing environment. Cloud RAN is considered a deployment option for future 5G networks to manage efficiency of implementation as described in section 3.2.4. A limited number of projects have indicated the deployment and use of Cloud RAN although many have indicated that Cloud RAN could be implemented but is not in scope.

Five projects explicitly use Cloud RAN with the selected functional split option. One project uses RAN processing in software albeit not in virtual machines. One project experiments with different functional splits, and massive MIMO. It virtualizes RAN functions with Docker containers.

The use of Cloud RAN is related to the previous deployment option on the functional split, hence there is no evident mapping of the implemented Cloud RAN to use cases or 5G service classes.

What transport technology do you use (if any)?

This question interrogates the use of different network transport technologies. The main network segment for transport is positioned between RAN/Edge and Core segments; however fronthaul, backhaul and other intermediate network transport technologies are included in the analysis. A number of projects rely on plain IP type transport such as UDP, TCP, MQTT over TCP, MPLS or various types of VPNs to interconnect testbeds and their components. Various SDN technologies are not explicitly listed, because these can be considered common. In certain cases specific technology options have been chosen for the implementation of the transport network segment as listed below:

- mmWave Mesh network at 26 GHz and mmWave backhaul at 60 GHz [6-6]
- Wireless backhaul IEEE 802.11ac [6-6]
- Analogue Radio over Fibre [6-11]
- Optical transport optionally with SDN [6-6] [6-10] [6-12]
- Passive WDM optics [6-6]
- Synchronous Ethernet [6-6]
- L2 Ethernet [6-6]
- L2SM (Layer 2 VPN Service Model) with Carrier Ethernet type service, and IP transport services over optical transport network [6-8]
- Satellite transport [6-9]
- Multilink backhaul including multipath TCP/QUIC [6-9]
- Next Generation Fronthaul interface (NGFI) and CPRI for RAN [6-12]
- LTE & eMBMS, Multi-link [6-13]

No evident mapping of the used transport technology to use cases or 5G service classes could be derived.

What spectrum do you use?

This question interrogates the use of spectrum either in existing LTE bands, unlicensed bands (WiFi) or 5G NR bands available for experimentation.

Table 6-4 Use of spectrum

Band / Frequency	Comment	Project reference
700 MHz Band 28	Dynamic Spectrum use	[6-14] [6-15]
1.8 GHz (LTE-FDD)	SDR	[6-16]
2.4 and 5.0 GHz WiFi		[6-17] [6-18] [6-16]
2.6 GHz (various LTE frequencies), LTE Band 3, LTE Band 7 (FDD), LTE Band 20, LTE Band 38 (TDD), LTE-A		[6-19] [6-20] [6-17] [6-18] [6-21] [6-14] [6-16] [6-22] [6-23]
~3GHz		[6-24]
3.5 GHz Band 42 (CBRS)		[6-20] [6-14]
3.5 GHz	5G NR in selected locations	[6-18] [6-21]
5 GHz	Wireless backhaul	[6-18]

Up to 6 GHz	SDR	[6-16]
Satellite K-bands (Ku 12-18 GHz), Ka (26-40 GHz)		[6-17] [6-25]
26 GHz, 40 GHz, 60 GHz mmWave	In building premises, backhauling, partly SDR, transport network	[6-26] [6-18] [6-16] [6-27]
C-Band, L-Band optical spectrum	Optical transport	[6-28]
Visible Light	In building premises	[6-26]

Concerning the use of spectrum, it can be noted that a limited number of projects uses spectrum specifically allocated to 5G NR at the 3.5 GHz and mmWave frequencies. Most of the projects rely on licensed LTE spectrum for use case experimentation. Satellite spectrum is used as allocated by the satellite operators.

Did you deploy SA/NSA options? (Stand-Alone/Non-Stand-Alone)

This question attempts to assess the readiness of the deployments towards a full 5G system.

Table 6-5 SA/NSA deployments

Deployment option	Project reference
Non-Stand-Alone	[6-19] [6-20] [6-29] [6-18] [6-25] [6-14] [6-22] [6-23]
Stand-Alone	None
Stand-Alone (planned)	[6-18] [6-25] [6-23]

It should be noted that due to the evolving standard for 5G core no complete Stand-Alone (SA) deployment was reported. In some cases this is planned in the last phase of certain projects. In many cases the deployments do not adhere to the NSA/SA deployment options.

Did you use EPC or 5G core?

This question attempts to assess the readiness of the deployments towards a full 5G system and is tightly couple to the previous question on the NSA/SA deployment options.

Table 6-6 EPC/5G Core deployments

Core Network	Project reference
5G Core	[6-24] [6-18] [6-25] [6-22]
EPC	[6-20] [6-21] [6-25] [6-14] [6-16] [6-22] [6-23] [6-15]

It should be noted that due to the evolving standard for 5G Core the deployed components should be mainly classified as prototypes and do fully support 5G Core specifications.

6.2 Management and orchestration

The substantial softwarisation of the 5G system requires increased automation of the management and orchestration functions that govern the interplay of the functional components in the overall system. As described in section 5, ETSI MANO emerges as one of the standard with related technologies that implement this automation layer. This sub-section analyses the extent to which the 5G PPP phase II projects used advanced management and orchestration technologies in the context of their use cases.

At which level do you use slicing and orchestration? (Resource level, Service level...)?

This questions attempts to interrogate the level of slicing and orchestration implemented in the use cases and at which level.

Table 6-7 Slicing and orchestration levels

Concept	Level	Responses	Comments
Slicing and orchestration	Service and resource	11	Extensions to OSM
	Service	5	
	Resource	2	
Slicing	Resource level	4	No RAN slicing (x2)
	Service level	1	A slice includes compute resources
Slicing (planned)	Resource	4	Slicing not well defined; exploring the concept. Focus on Transport segment
Orchestration	Service	4	
No slicing and orchestration		1	

From the analysis of the responses it can be seen that a majority of the deployed use cases use slicing and orchestration at both resource and service levels to achieve some level of automation (see next question). A smaller number of projects indicate the use of slicing and orchestration at either the service or the resource level only.

Concerning the application of the slicing concept only, two areas of improvements can be identified; namely that the slicing concept is not well defined or standardised yet and that in at least two cases the RAN segment has been excepted from slicing. Furthermore the plans to still introduce slicing in the deployed use cases mainly focus on the transport network segment, which may indicate that this segment receives priority in the application of the concepts. One project indicates no use of slicing and orchestration concepts.

Do you orchestrate end-to-end? If not, which parts?

Connected to the previous question, yet orthogonal to it, this question interrogates the extent of orchestration use in the deployed use cases.

End-to-end	Responses
No	6
Yes	12
Partly	Except PHY/MAC, among VNFs (x2), Vertical Application, RAN and Core, Fronthaul

About half of the deployed use cases are enabled for end-to-end orchestration at the level at which they are enabled. About a quarter of the use cases use orchestration at a limited scope.

What statements can you make with respect the level of automation that you achieved as a function of Scalability, Dynamicity, Number of instances ...

This question interrogates the level of automation achieved with respect to the main properties subject to automation.

Table 6-8 Level of automation³

	Scalability	Dynamicity	Number of instances
Manual	6	6	10
Automation	13	16	9
No	4	2	4

Considering the level of automation a remarkable high number of use case deployments achieved automation levels with respect to the properties Scalability, Dynamicity, Number of instances etc. Further aspects auf automation include “*service composition*” and “*service federation*” that were indicated in addition.

It can be noted that certain use case deployments have been accepted as ETSI ENI PoCs (Proof of Concept) and that for certain use cases the automation properties meet pre-commercial deployment requirements.

Do you employ any kind of autonomous capability (closed loop control/management)?

This question interrogates the application of autonomous properties in the use case deployments. Autonomic properties are understood in the context of the use of cognitive algorithms or existing concepts from autonomic computing such as MAPE (Monitor-Analyse-Plan-Execute). In contract autonomic properties (see previous question) refer usually to the application of policy based execution of management tasks to enforce service level agreement boundaries. The solutions that have been introduced include:

- MAPE loop focused on optimization of media services
- Autonomous wavelength control in passive WDM, with zero touch provisioning support
- Semi-autonomous traffic type detection and prioritisation
- Early failure detection using monitoring, data analytics and autonomic closed loop management and control at network subsystem level
- Autonomous SLA management based on configuration, monitoring and reaction to alerts

³ The figures in the table indicate responses related to use cases

- Resource assignment utilizing traffic prediction models.
- Cognitive network management, enabling machine-learning-empowered autonomous control loops for slice FCAPS management and slicing control for pro-active failure detection and handover prediction.

6.3 Verification and Validation System

In 5G, to reduce the time-to-market for networked services and to lower the entry barrier to third party developers of Virtual Network Functions (VNFs) and Network Services (NSs), an integrated Development and Operations (DevOps) methodology is crucial. One of the biggest challenges in DevOps is the Validation and Verification (V&V) of individual VNFs and NSs so that providers of these services can be sure of their behaviour. Such a V&V process does not only include functional testing of VNFs and NSs but also non-functional tests, such as performance measurements for gaining insights about resource requirements to fulfil SLAs and to provide the expected Quality of Experience (QoE). To fit seamlessly into the anticipated DevOps workflow, all these V&V procedures need to be fully automated and be able to qualify any VNF or NS without further human interaction.

The future network service ecosystem is expected to have various stakeholders who are taking on roles in validating and verifying network services. It is foreseen that the developer would want to ensure the quality of their own code prior to release, and the network service provider would want to validate and verify all code prior to deployment on their network. Finally, it is expected that a range of third-party verification and validation organisations would exist that would amortise the cost of verification and validation for independent network service and VNF developers who hope to supply many network service providers.

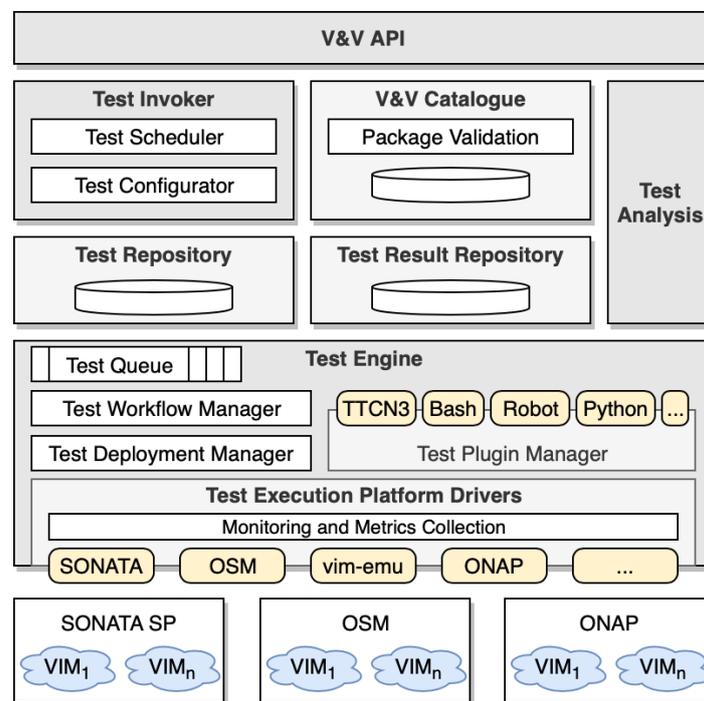


Figure 6-2: Verification & Validation Platform Architecture

Figure 6-2 shows the internal architecture of a V&V platform and its surrounding building blocks. It is highly modularised and consists of the following main components that enable a fully automated V&V workflow: (i) The V&V API, allowing V&V platform users to submit packages for verification and validation; (ii) the Test Invoker, responsible for the test case configuration,

scheduling, and maintenance of the test state; (iii) the V&V Catalogues holding the artefacts to be tested, e.g., VNFs and network services; multiple repositories, i.e. the Test Repository and the Test Result Repository, are used to store tests, test results, as well as raw monitoring metrics collected during the tests; (iv) the Test Engine responsible to control the execution of tests in the test queue using an extensible set of test plugins. The V&V platform uses the concept of plugable Test Execution Platform Drivers to abstract and unify the interface towards the test execution platforms on which the VNFs or services under test (SUT) are deployed and the tests are actually executed. Finally, there is a set of tools for Test Analysis. It is important to note that the MANO system is part of the system under test, as it impacts the performance of VNFs and NSs, in particular in relation to scaling and fail-over.

Different types of tests can be distinguished: functional, performance, syntax, API, and security. The goal of categorizing the tests is to facilitate the test suite developer or the V&V provider to efficiently search and find the test they need to test a system comprehensively, as well as to facilitate to define priority of tests if the testing resource is limited.

The results of tests executed on the V&V testing platform need to be managed to facilitate the further use and analytics of the data. In addition to actual measurements, test data needs include information on the test plan, the test profile, and the test environment configuration. Test results are stored in a dedicated repository that enables other modules and end-users (e.g. developers) to manage the test results.

6.4 Emulation Framework

The emulator component provides a light-weight local rapid-prototyping environment for debugging services and components. It was created to support network service developers to locally prototype and test their network services in realistic end-to-end multi-PoP scenarios. It allows the execution of real network functions, packaged as Docker containers, in emulated network topologies running locally on the developer's machine. The emulation platform also offers OpenStack-like APIs for each emulated PoP so that it can integrate with MANO solutions, like OSM.

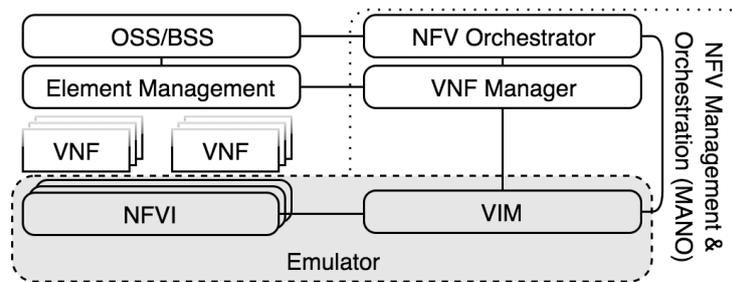


Figure 6-3: Scope of the emulation platform in the simplified ETSI NFV reference architecture

Figure 6-3 shows the scope of the emulator and its mapping to a simplified ETSI NFV reference architecture in which it replaces the network function virtualisation infrastructure (NFVI) and the virtualised infrastructure manager (VIM). The emulator allows the use of standard Docker containers as VNFs within the emulated network. It automatically starts OpenStack-like control interfaces for each of the emulated PoPs which allow MANO systems to start, stop and manage VNFs. Specifically, the emulator provides the core functionalities of OpenStack's Nova, Heat, Keystone, Glance, and Neutron APIs. Even though not all of these APIs are directly required to manage VNFs, all of them are needed to let the MANO systems believe that it manages a real-world multi-VIM deployment, i.e., the MANO system's southbound interfaces can connect to the OpenStack-like VIM interfaces of each emulated PoP.

The emulation platform is designed to act as a local test execution platform that can be installed locally on a developer's laptop. The benefit of such a local platform are the quick turnaround times for the developer who tries to fix bugs inside a network service. At the same time, test developers benefit from the availability of an easily accessible test environment. However, due to the nature of such an emulation environment, it mostly focuses on functional tests rather than on performance tests. In addition to serving as a test environment for VNFs and NSs, the emulation platform can utilize its ability to emulate VIM interfaces to be used to test different MANO and service platform solutions.

6.5 References

- [6-1] 5G PPP contractual arrangement, online <https://5g-ppp.eu/contract/>
- [6-2] 5G empowering vertical industries, brochure available online at https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf
- [6-3] 3GPP TS 28.554, Technical Specification Group Services and System Aspects; Management and orchestration; 5G end to end Key Performance Indicators (KPI)
- [6-4] ITU-T Recommendation E.800: "Definitions of terms related to quality of service".
- [6-5] Definition of the Testing Framework for the NGMN 5G pre-commercial Network Trials, NGMN Alliance, January 2018, Version 1.0, available online https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2018/180220_NG_MN_PreCommTrials_Framework_definition_v1_0.pdf
- [6-6] 5G-PICTURE deliverable D2.2, https://www.5g-picture-project.eu/download/5g-picture_D2.2.pdf
- [6-7] 5G-MoNArch deliverable D2.3 https://5g-monarch.eu/wp-content/uploads/2019/05/5G-MoNArch_761445_D2.3_Final_overall_architecture_v1.0.pdf
- [6-8] METRO-HAUL deliverable D3.1 <https://zenodo.org/record/2586698>
- [6-9] SAT5G deliverable D3.2 https://www.sat5g-project.eu/wp-content/uploads/2019/04/SaT5G_D3.2_ADS_v2.0_D.pdf
- [6-10] 5G-TRANSFORMER deliverable 1.3 http://5g-transformer.eu/wp-content/uploads/2019/05/D1.3_5G-TRANSFORMER_Refined_Architecture.pdf
- [6-11] 5G-PHOS deliverable D2.1 http://www.5g-phos.eu/pdf/5G-PHOS_D2.1_Final.pdf
- [6-12] SLICENET deliverable D2.2 https://doi.org/10.18153/SLIC-761913-D2_2
- [6-13] 5G-XCAST deliverable D4.2 http://5g-xcast.eu/wp-content/uploads/2018/10/5G-Xcast_D4.2_v2.0_web.pdf
- [6-14] MATILDA project <http://www.matilda-5g.eu/>
- [6-15] 5G-XCAST project <https://5g-xcast.eu/>
- [6-16] 5G-TRANSFORMER project <http://5g-transformer.eu/>
- [6-17] NRG5 project <http://www.nrg5.eu/>
- [6-18] 5G-PICTURE project <https://www.5g-picture-project.eu/>
- [6-19] 5G ESSENCE project <http://www.5g-essence-h2020.eu/>
- [6-20] 5GCITY project <https://www.5gcity.eu/>
- [6-21] 5GCAR project <https://5gcar.eu/>
- [6-22] NGPaaS project <http://ngpaas.eu/>
- [6-23] SliceNet project <https://slicenet.eu/>
- [6-24] 5G-MONARCH project <https://5g-monarch.eu/>
- [6-25] SAT5G project <https://www.sat5g-project.eu/>
- [6-26] IoRL project <https://iorl.5g-ppp.eu/>
- [6-27] 5G-PHOS <http://www.5g-phos.eu/>
- [6-28] METRO-HAUL project <https://metro-haul.eu/>
- [6-29] 5G-MEDIA project <http://www.5gmedia.eu/>

7 Impact on standardization

The contributions to standards that shape the development of 5G is among the high impact outcomes expected from 5G PPP projects. This chapter surveys the contributions to standards made by different 5G PPP Phase 2 projects in areas associated with the development of 5G architectural concepts as outlined previously in this report.

7.1 Brief profile of targeted SDOs

The standards development organizations (SDOs) that have been targeted by the 5G PPP Phase 2 projects in the scope of this report are profiled briefly in this subsection. Particular focus is placed on the major SDOs, while also shining a spotlight on other smaller ones and specialised industry alliances with specific interests in technology specifications that also impact 5G architecture developments.

7.1.1 3GPP

The Third-Generation Partnership Project (3GPP) has been primarily responsible for specifying and maintaining standards of current and preceding generations of mobile communications technologies. The standards development work in 3GPP is organized in Technical Specification Groups (TSGs), namely: Radio Access Networks (RAN), Service & Systems Aspects (SA) and Core Network & Terminals (CT). Each of these TSGs further constitute multiple Working Groups (WGs). Furthermore, 3GPP adopts the three stage phased approach for specifying standards with Stage 1, Stage 2 and Stage 3 focusing on service requirements, architecture and detailed interfaces specifications, respectively.

The 3GPP SA WG2 – Architecture covers aspects related to Architecture with a system-wide view, deciding on how new functions integrate with existing network entities. Its main responsibility is developing Stage 2 of the 3GPP network. It picks up on work within SA WG1 – Service Requirements, identifying the main functions and entities of the network, how these entities are linked to each other and the information they exchange. SA WG2 outputs are used as inputs for groups in charge of defining the precise format of messages in Stage 3.

In addition to SA WG2, SA WG4 (Codec) also contributes to the overall architecture. Part of its mandate includes quality evaluation, end-to-end performance, and interoperability aspects with existing mobile and fixed networks, from a Codec perspective.

TSG RAN is responsible for Stage 2 specification of the Radio Access Network. Radio and edge architecture falls under three RAN WGs: RAN WG1 – Radio Layer 1; RAN WG2 – Radio Layer 2 and Radio Layer 3 RR; RAN WG3, responsible for the overall UTRAN/E-UTRAN architecture and protocol specifications. RAN WG1 focuses on the physical layer of the radio interface, while RAN WG2 is in charge of the Radio Interface architecture and protocols.

Core and transport architecture falls under the previously discussed SA WG2 and CT WG1 which is responsible for specifications that define the User Equipment (UE) – core network Layer 3 radio protocols and core network side of the lu reference point; CT WG4, dealing with the Bearer Independent Architecture, among other aspects.

The MANO aspects falls to SA WG5 – Network Management, which specifies the requirements, architecture and solutions for provisioning and management of the network (RAN, CN, IMS) and its services.

7.1.2 ETSI

The European Telecommunications Standards Institute (ETSI) has standardization activities that specify requirements and potential enablers or building blocks for an overall 5G system. This includes key aspects, such as, NFV, MEC and MANO, which have a significant impact on the 5G architectural developments. A number of ETSI Industry Specification Groups (ISGs), Technical Committees (TCs) and projects have active collaboration with 3GPP and, in some cases, provide direct input to 3GPP [7-1]. Some of these ETSI entities that are relevant in the scope of this report are briefly reviewed below.

- ETSI Zero Touch network and Service Management (ZSM) is an ETSI ISG that specifies horizontal (i.e., cross-domains, cross-technology) and vertical (i.e., cross layers) end-to-end network and service management reference architecture to enable agile, efficient and qualitative management and full automation of emerging and future networks and services [7-2]. Full automation in this context includes automation of delivery, deployment, configuration, assurance and optimization of networks and services.
- ETSI NFV ISG defines requirements and architecture for the virtualization of network functions, as well as, addressing technical challenges of network virtualization [7-3]. The outputs of this ISG group includes pre-standardization studies, detailed specifications, and Proof of Concepts.
- ETSI Open Source MANO (OSM) is an ETSI-hosted project that focuses on providing an open source NFV Management and Orchestration (MANO) software stack aligned with ETSI NFV Information Models [7-4].
- ETSI TC on Satellite Earth Stations and Systems (SES) focuses on all aspects related to satellite earth stations and systems [7-5]. This includes satellite communication systems, services and applications; as well as, satellite navigation systems and services; all types of earth stations and earth station equipment.
- ETSI MEC ISG aims for a standardized, open environment that will enable the efficient and seamless integration of applications from vendors, service providers, and third-parties across multi-vendor MEC platforms [7-6].
- ETSI Experiential Networked Intelligence (ENI) ISG is currently defining a Cognitive Network Management architecture using closed-loop AI mechanisms that leverage context-aware and metadata-driven policies to improve the operator experience [7-7].

Additionally, ETSI organizes Plugtests to ensure interoperability of products and services, as well as provide feedback to different standardization groups (both within and outside ETSI) [7-8]. For instance, the ETSI Mission Critical Push To Talk (MCPTT) Plugtests events⁴ have provided a platform to demonstrate the interoperability of a wide range of implementations utilizing different scenarios and test cases based on 3GPP Mission Critical Services.

7.1.3 ITU

The International Telecommunication Union (ITU) coordinates the development of global telecommunications standards in addition to fostering the growth and sustained development of the sector and ensuring universal access. The activities of ITU activities are focused on three core sectors, namely: Standardization (ITU-T) which standardizes global telecommunications; (2) Radiocommunications (ITU-R) sector, which manages the international RF spectrum and satellite orbit resources, and Development (ITU-D) that supports the ITU mission to ensure equitable, sustainable and affordable access to ICT. The ITU sectors include Study Groups (SGs), which assemble global experts for the development of international standards commonly referred to as

⁴ The Plugtest event name was changed from MCPTT to MCX (Mission Critical Services) in order to increase the event scope as vendors are moving beyond voice.

ITU-x Recommendations where x stands for T, D or R depending on the ITU sector concerned. In terms of the scope of this report, SGs of interest include:

- ITU-T SG13 - Future networks, which focuses on IMT-2020 (that defined the requirements for 5G networks and services), cloud computing and trusted network infrastructure [7-9].
- ITU-T SG15 - Networks, Technologies and Infrastructures for Transport, Access and Home [7-10]. This SG gives special consideration to the changing telecommunication environment towards future networks, including networks that are supporting the evolving needs of mobile communications (IMT-2020).

7.1.4 IETF

The Internet Engineering Task Force (IETF) is an open SDO in the area of Internet-related technologies. In the context of 5G, the main areas that IETF is focusing on includes network slicing, MEC, machine learning at network level, and Low Power IoT Networking (LPWA). The technical work in IETF is performed in Working Groups (WGs), which are organized based on topic into several technical areas. The IETF standards produced are then published as Internet Drafts which may evolve into accepted Request for Comment (RFC) documents [7-11].

The Common Control and Measurement Plane (CCAMP) WG is responsible for standardizing a common control plane and a separate common measurement plane for non-packet technologies (e.g. optical cross-connects, microwave links, TDM switches etc.) found in the Internet and in the telecom service provider networks [7-12].

Furthermore, within the IETF framework there is a possibility to organise pre-WG technical discussions in the form of Birds of a Feather (BoF) sessions at IETF meetings [7-13]. While some of the BoF may eventually evolve into fully-fledged WGs, others simply provide a discussion forum on topics of possible interest within the IETF community. Common Operations and Management on network Slices (coms) is one example of a BoF that appears later among the standards contributions of this report [7-14].

7.1.5 Other SDOs

In addition to the major SDOs mentioned previously, the 5G PPP Phase 2 projects also have targeted SDOs or technical specification groups that are based in industry alliances (representing interests of particular industry groups) and open source projects that adopt open source principles for producing standards or other open specifications. The groups that appear in the reported contributions in this report include:

- DVB (Digital Video Broadcasting) is an industry-led consortium of the world's leading digital TV and technology companies that develops open technical specifications for the delivery of digital TV and other broadcast services. These DVB specifications are ultimately translated into international standards by major SDOs, such as, ETSI. The specification work is carried out in DVB WGs, such as, DVB TM-IPI which is responsible for the development of technical specifications for the delivery and discovery of DVB services over IP networks [7-15].
- The Open Networking Foundation (ONF) is a user-driven non-profit organization focused on promoting the adoption of SDN through open standards development [7-16].
- MulteFire Alliance specifies an LTE-based technology for operating in unlicensed & shared spectrum [7-17].
- MEF (formerly known as the Metro Ethernet Forum) is an industry alliance specifying agile, assured, and orchestrated communications services across a global ecosystem of automated networks [7-18].

- The Open ROADM Multi-Source Agreement (MSA) defines interoperability specifications for Reconfigurable Optical Add/Drop Multiplexers (ROADM) [7-19]. The specifications consist of both Optical interoperability as well as YANG data models.
- The NGMN Alliance in an industry alliance that complements and supports SDOs (e.g. 3GPP) by providing a coherent view of requirements of mobile operators for next generation networks with a particular focus on 5G. The specification work is initiated through a number of projects. For instance, the NGMN Network Management and Orchestration (NWMO) project has been specifying the requirements for 5G Network and Service Management including Orchestration [7-19].

7.2 Standards impact by 5G PPP Phase 2 projects

This subsection summarizes details of over 200 standards contributions from 5G PPP Phase 2 projects that directly or indirectly influence 5G architectural developments. The standards contributions reported come in many formats including technical or specification documents, presentations, white papers, proof of concepts, interoperability tests, source code and so on. For each contribution, details are provided of the contributing project, SDO targeted, title or short description of the contributions and status of the contribution at the time of writing. Furthermore, the listing of the contributions is mapped to relevant subsections with each subsection representing the architectural areas addressed in this report. The breakdown of the contributions across the four areas is shown in Table 7-1. A complete listing of these contributions is provided in the Annex of Chapter 0 and briefly analysed below. In each case, we provide statistics on the highest concentration to the target SDO.

Table 7-1: Number of contributions reported by 5G PPP Phase 2 project per architectural area

Number of contributions per architectural area	
Overall architecture	70
Radio and edge architecture	41
Core and transport architecture	58
Management and orchestration architecture	50
Total	219

7.2.1 Contributions related to overall architectures

The contributions by 5G PPP Phase 2 projects related to the 5G overall architectures have mostly targeted 3GPP TSG SA. A significant number of those contributions are related to implementation 5G V2X systems and multimedia broadcast or streaming services.

The bullet points below show the breakdown for 3GPP SA Working groups, with a high concentration to SA2, followed by SA4.

- 3GPP SA2 – Architecture: 40 contributions.
- 3GPP SA4 – Codec: 25 contributions.
- 3GPP SA6 – Mission-critical applications: 3 contributions
- 3GPP SA1 – Services: 1 contribution
- 3GPP SA5 – Telecom Management: 1 contribution.

The figure below shows the overall concentration of inputs for overall architecture.

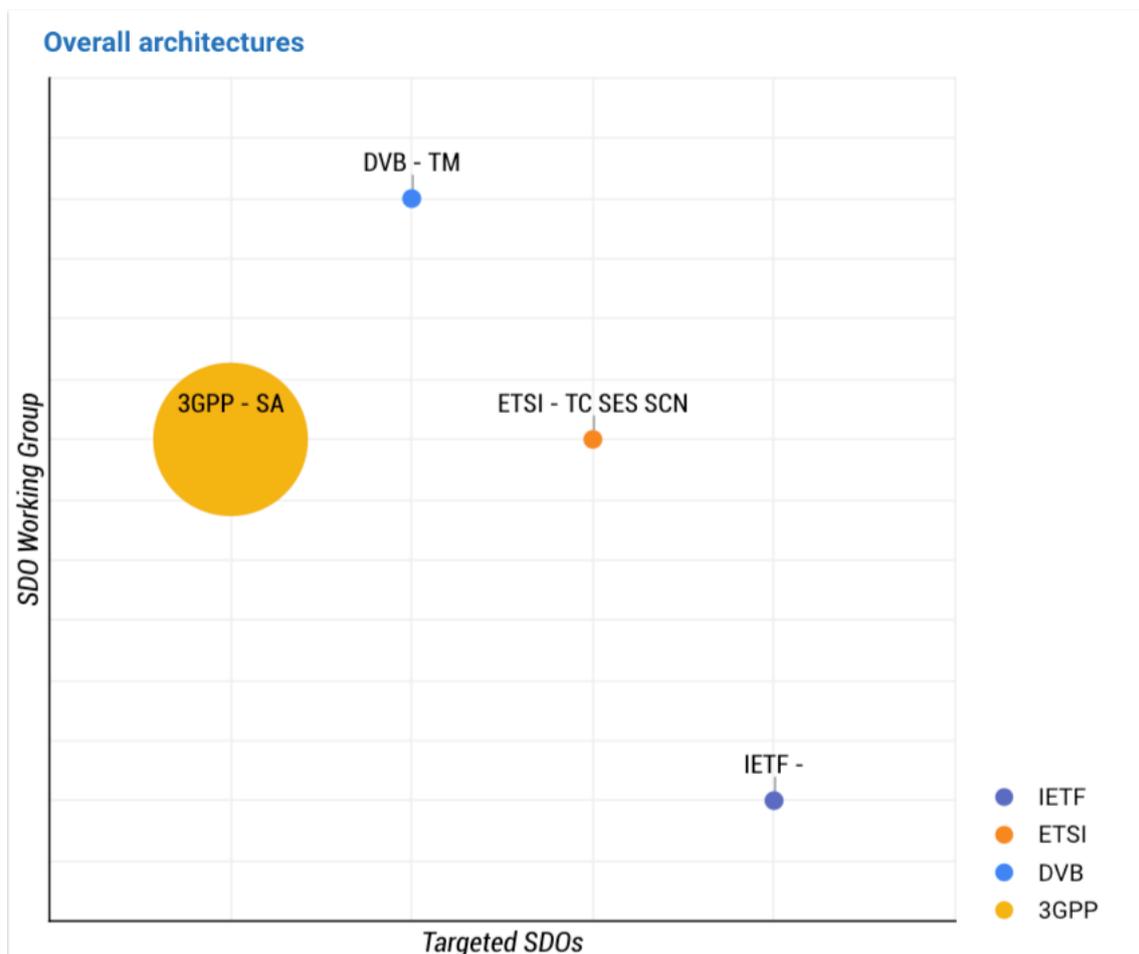


Figure 7-1: SDOs targeted for contributions related to overall architectures

7.2.2 Contributions related to radio and edge architectures

The 5G PPP Phase 2 project contributions related to RAN architectures were mostly to WGs of the 3GPP TSG RAN (see Figure 7-2). To that end, these contributions focused on 5G NR enhancements for V2X and multimedia broadcast. Additional standards contributions for the multimedia broadcast are also targeted towards the DVB industry alliance [7-15]. In the case of edge architectures, ETSI (MEC and NFV ISGs) have been for main venues for MEC-related contributions primarily addressing 5G architectural enablers for MEC applications.

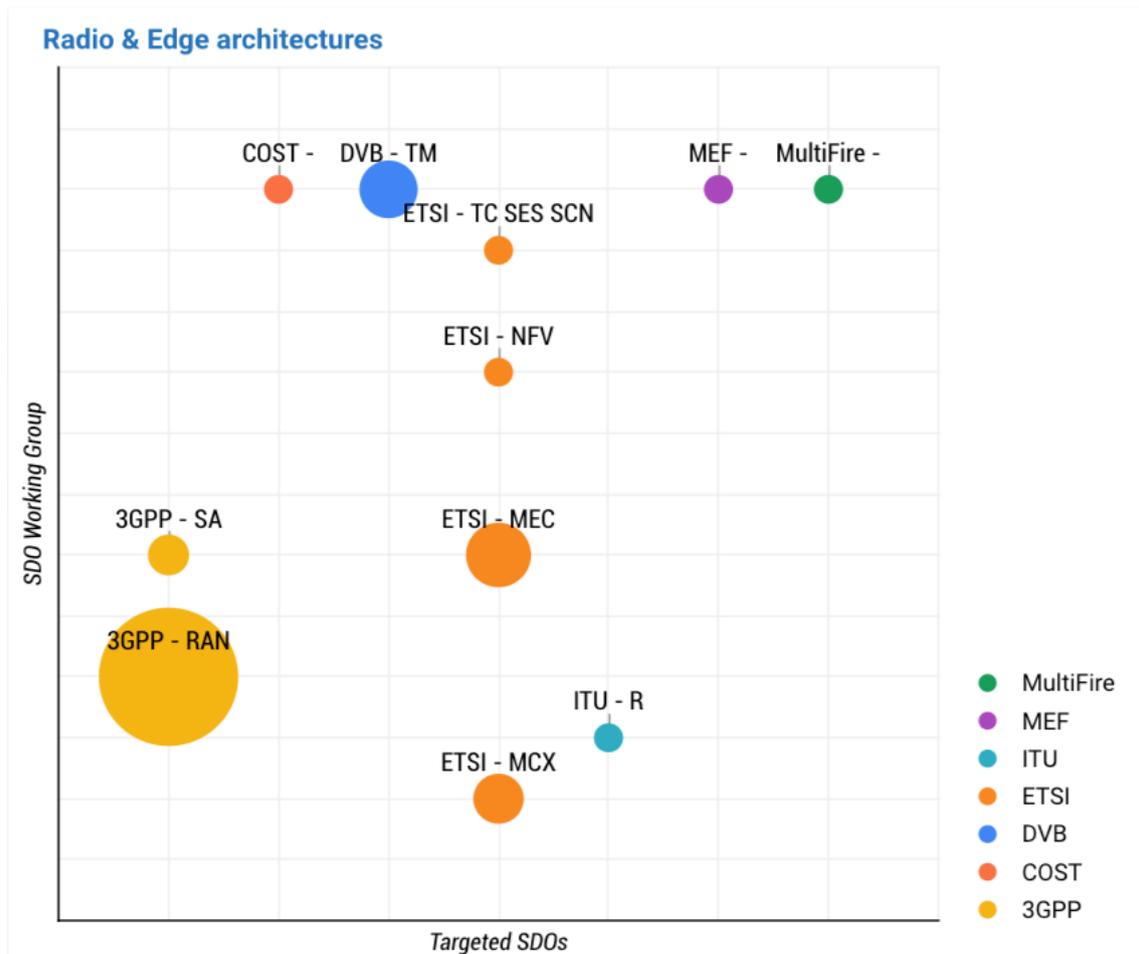


Figure 7-2: SDOs targeted for contributions related to radio and edge architectures

The bullet points below show the breakdown for 3GPP RAN:

- 3GPP RAN1– Radio layer 1: 10 contributions.
- 3GPP RAN2– Radio layers 2 and 3: 7 contributions.
- 3GPP RAN3– UTRAN/E-UTRAN architecture: 4 contributions.

7.2.3 Contributions related to core and transport architectures

The contributions for 5G core network architectures have mostly been targeted towards WGs of 3GPP TSGs SA and CT (see Figure 7-3). As for contributions related to transport architectures, those for microwave/millimetre wave transport have been towards IETF, whereas, those contributions for optical-based transport have targeted mostly ITU-T and IETF.

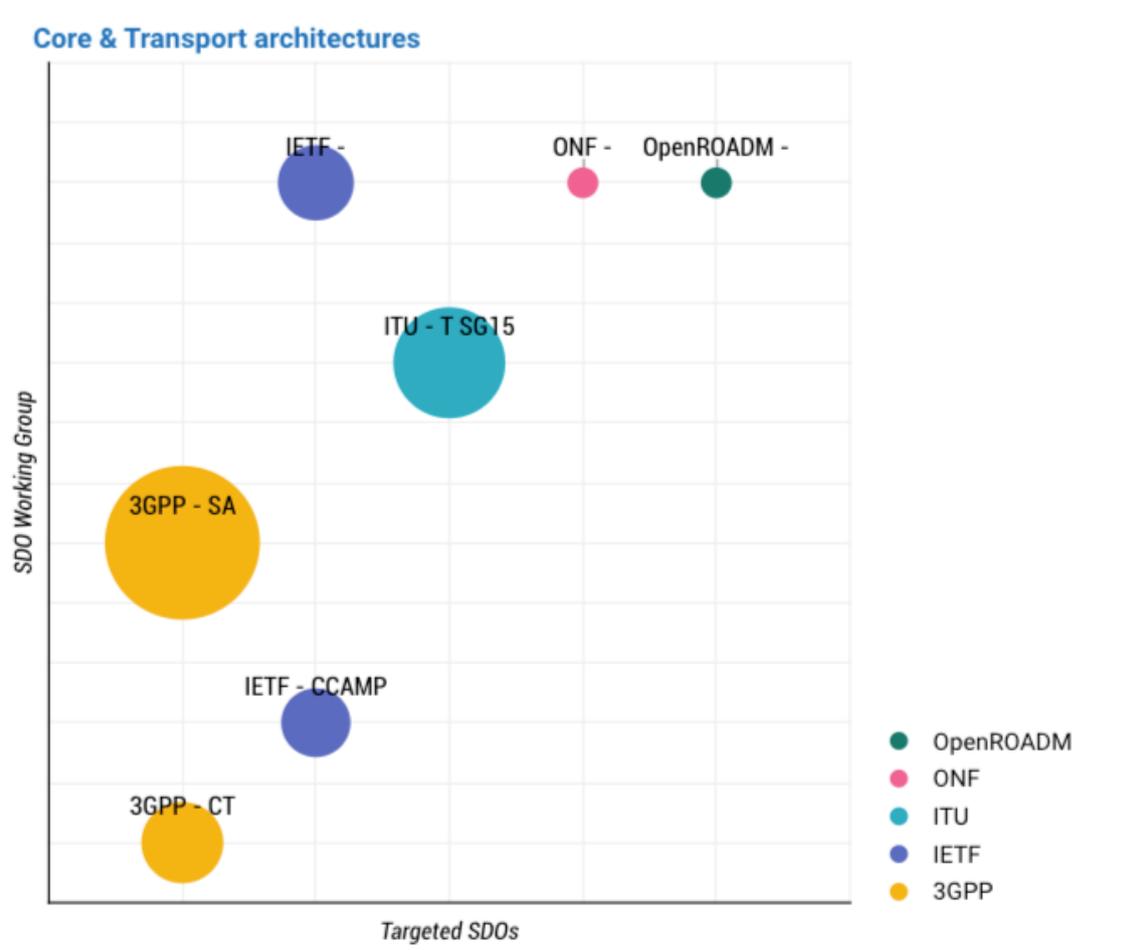


Figure 7-3: SDOs targeted for contributions related to core and transport architectures

The bullet points below show the breakdown for 3GPP SA and CT.

- 3GPP CT1– Architecture: 8 contributions.
- 3GPP CT4– Architecture: 11 contributions.
- 3GPP SA2– Architecture: 21 contributions.

7.2.4 Contributions related to management and orchestration architectures

The contributions of 5G PPP Phase 2 projects related to MANO have been mostly to ETSI via the ZSM ISG, but also to the NFV ISG and OSM project (see Figure 7-4). Outside of ETSI, there have been contributions to 3GPP mostly targeting the SA WG5 (Telecom Management), which specifies architecture and solutions for provisioning, charging and management of mobile networks (including RAN and core) and their services.

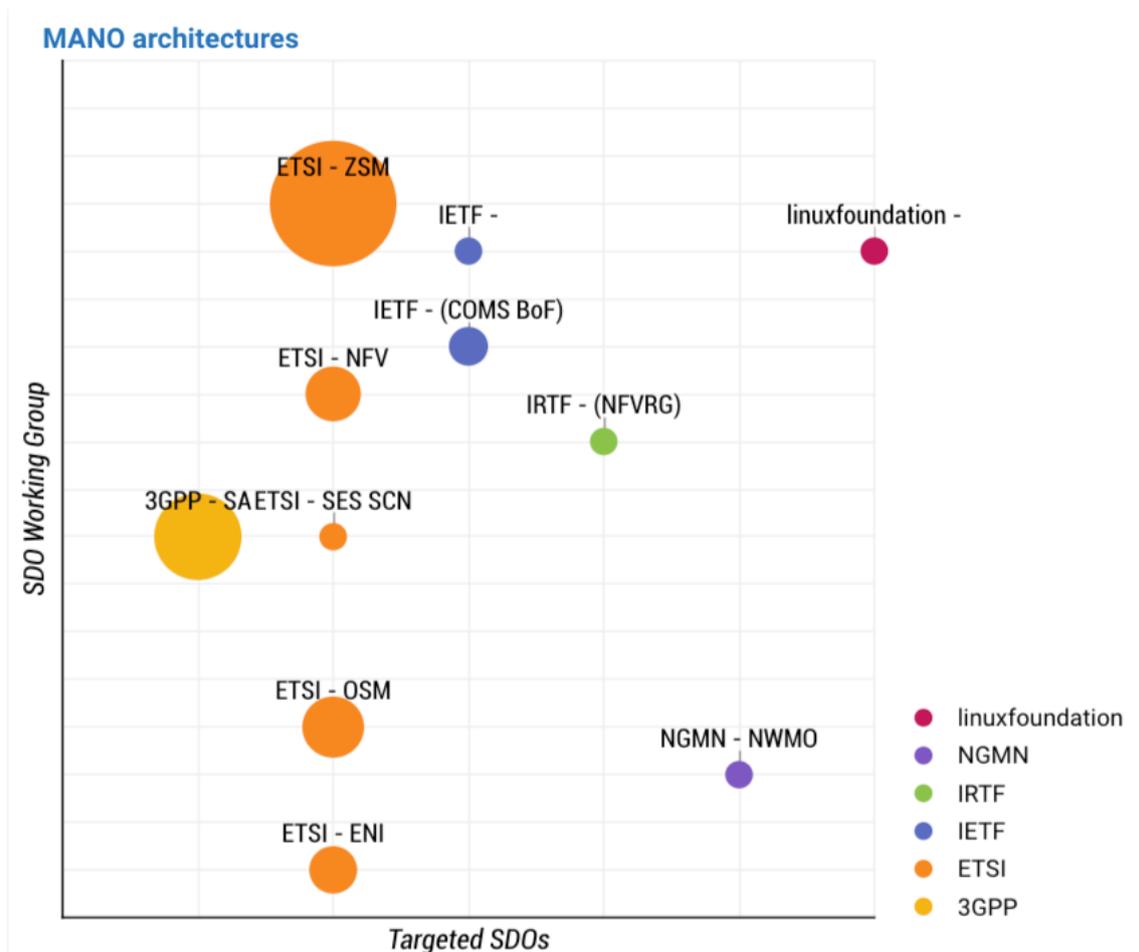


Figure 7-4: SDOs targeted for contributions related to management and orchestration architectures

7.3 Fostering Standardization via 5G PPP

The 5G Infrastructure Association (5G-IA) Pre-Standardization WG (in parallel with the 5G PPP Architecture WG) continues to track contributions to 5G standardization, including open source initiatives. It also monitors study and work items across relevant SDOs. Its primary focus is on 3GPP. For example, tracking the status of study and work items for Release 17 (March 2020), sharing updates on EU priorities, and working on a gap analysis for Release 18 and beyond. In addition to the Phase 2 projects, the WG has already on-boarded specialists from Phase 3 projects (initially projects from ICT-17 calls and then project ICT-18 and ICT-19 and 18, and will be supporting the smooth participation of ICT-19 from June 2019.

Besides this, a special Task Force has been set up on 5G Standardization and Vertical Industries. The Task Force comprises key members of the 5G-IA and its Working Groups:- 5G-IA Chairman; Chair of the Verticals Task Force; Chair of the Trials Working Group and Activity Leader of the Pre-Standardization Working Group. It also comprises high-profile members of 3GPP, several of its Market Representation Partners including the 5G-IA, 5G Automotive Association (5GAA), Public Safety Communications Europe (PSCE) and 5G Alliance for Connected Industries and Automation (5G-ACIA), as well as high-profile ETSI representatives. This Task Force works to support and encourage vertical industries in contributing to the 5G standardization process, including a workshop series and practical guides. Outputs from this work will feed into plans for inputs and support of common requirements and complementarities. They will also help define the future involvement of other relevant SDOs so standardization work is complementary and globally harmonized.

7.4 References

- [7-1] Sampson N. (2017). ETSI views and contribution on 5G standardisation. ETSI Board 5GCOM#18, May 2017.
- [7-2] ETSI. Zero touch network & Service Management (ZSM) ISG. URL: <https://www.etsi.org/technologies/zero-touch-network-service-management>.
- [7-3] ETSI. Network Functions Virtualization (NFV) ISG. URL: <https://www.etsi.org/technologies/nfv>.
- [7-4] ETSI. Open Source MANO (OSM). URL: <https://osm.etsi.org>
- [7-5] ETSI. TC on Satellite Earth Stations and Systems (SES). URL: <https://www.etsi.org/committee/1412-ses>
- [7-6] ETSI. Multi-access Edge Computing (MEC) ISG. URL: <https://www.etsi.org/technologies/multi-access-edge-computing>.
- [7-7] ETSI. Experiential Networked Intelligence (ENI) ISG. URL: <https://www.etsi.org/technologies/experiential-networked-intelligence>.
- [7-8] ETSI. ETSI Plugtest events. URL: <https://www.etsi.org/events/plugtests>
- [7-9] ITU-T. SG13 - Future networks. URL: <https://www.itu.int/en/ITU-T/about/groups/Pages/sg13.aspx>
- [7-10] ITU-T. SG15 - Networks, Technologies and Infrastructures for Transport, Access and Home. URL: <https://www.itu.int/en/ITU-T/about/groups/Pages/sg15.aspx>
- [7-11] IETF. (1996). RFC 2026 “The Internet Standards Process -- Revision 3,” October 1996 (updated March 2013). URL: <https://datatracker.ietf.org/doc/rfc2026/>
- [7-12] IETF. Common Control and Measurement Plane (CCAMP) WG. URL: <https://datatracker.ietf.org/wg/ccamp/about/>
- [7-13] IETF. Birds of a Feather (BoF) sessions. URL: <https://www.ietf.org/how/bofs/>
- [7-14] IETF. Common Operations and Management on network Slices (coms). URL: <https://datatracker.ietf.org/wg/coms/about/>
- [7-15] DVB. TM-IPI WG. URL: <https://www.dvb.org/groups/TM-IPI>

- [7-16] ONF. Open Networking Foundation (ONF). URL: <https://www.opennetworking.org/>
- [7-17] Multefire. Multefire alliance. URL: <https://www.multefire.org/>
- [7-18] MEF. MEF industry alliance. URL: <http://www.mef.net/>
- [7-19] OpenROADM. Open ROADM Multi-Source Agreement (MSA). URL: <http://openroadm.org/home.html>
- [7-20] NGMN. (2019). RFC 2026 “5G Network and Service Management Including Orchestration,” March 2019. URL: https://www.ngmn.org/fileadmin/ngmn/content/downloads/Technical/2019/190312_5G_Network_and_Service_Management_including_Orchestration_3.14.0.pdf

8 Conclusions and Outlook

5G is the first generation of mobile networks that is by design addressing the wide range of needs of the vertical industries. 5G offers unlimited mobile broadband experience, provides massive connectivity for everything from human-held smart devices to sensors and machines, and most importantly, it has the ability to support critical machine communications with instant action and ultra-high reliability. First 5G specifications are already available with 3GPP Rel.15; however, further enhancements and optimizations are needed to design a 5G System that meets the requirements from the vertical industries.

On this basis, 5G architecture enables new business opportunities meeting the requirements of large variety of use cases as well as enables 5G to be future proof by means of (i) **enabling E2E network slicing**, (ii) **addressing both end user and operational services**, (iii) **supporting softwarization and programmability natively**, and (v) **integrating novel NR technologies** (including fixed and wireless technologies).

Capitalizing on the vision and the requirements provided in the first two versions of the white paper, this white paper has highlighted the consolidated outcome from 5G PPP Phase 2 and Phase 3 projects. The main findings of this white paper are outlined in the following.

The overall architecture in Chapter 2 provides several enhancements to the 3GPP Rel.15 system architecture to address specific requirements from vertical industries. The introduction of management domains separates resources or functions according to technological or administrative criteria and provides the interfaces to seamlessly compose and manage e2e communication services that are built from resources of one or several domains. Moreover, network programmability, such as, data plane programmability, facilitates more dynamic and flexible customization of NFs and underlying resources. Further, the structure of the overall architecture supports the 5G ecosystem stakeholder model of the 5G PPP.

In Chapter 3, the description of the RAN and edge architecture is provided, focusing on the protocol extensions designed to support the wide and diverse variety of requirements introduced by verticals. Said variety makes it impossible for one single solution to support all of them, which is reflected on the solution presented, including network slicing for supporting QoS differentiation, as well as solutions such as virtual small cells, MEC enhancements, and local end-to-end paths, each designed to satisfy different typologies of requirements. Furthermore, different access technologies are considered, including 3GPP's, Wi-Fi, and visible light communication. The chapter provides an overview of the radio and edge technologies that will be available to satisfy the specific needs of the novel applications which are expected to arise in the upcoming decade.

Chapter 4 discussed the enhancements to the core network architecture and advanced transport technologies that are deemed necessary for the evolution of 5G networks to truly meet the requirements from vertical industries. The first 5G specifications do not include any multicast and broadcast capabilities. Building on the design principles of 5G, multicast can be introduced as part of connectivity service by enhancing functionalities of existing NFs and interfaces on one hand. On the other hand, new NFs can be introduced to the core network architecture to offer multicast and broadcast as a service via well-defined API to support terrestrial broadcast and other various vertical use cases. 3GPP SA approved a study item for SA2 working group on architectural enhancements for 5G multicast-broadcast services for Rel. 17. It can be expected that Rel.17 will be the first release of 5G specification that will support at least some multicast and broadcast capabilities. The 5G network is designed to meet heterogeneous requirements, which also means that the operation of 5G networks with many slice instances will become very complex. A data analytics framework becomes a necessity for successful operation of the system. The presented analytics framework introduces data analytics functions to the mobile system architecture (core network, RAN, data network, application function level, etc.) that offer data collection capabilities and data analytics functionalities. The analytics framework is a key enabler

for network monitoring, analysis, optimization, assessment and assurance. The 3GPP system architecture in Rel.15 includes network data analytics function however its functionalities are limited to the provisioning of load level information per network slice instance. In O-RAN, near-real time operations have been defined to capture operations like QoS management, traffic steering, mobility management, etc. A demand for performance data and analytics may increase as the number of commercial 5G network deployments grows, which may consequently translate to a need for further standardization of data analytics functions and interfaces in the system as described in the analytics framework.

On the other hand, the requirement to provide infrastructure connectivity from the APs to the CN is provided through the *transport network*, interconnecting NFs, CN and RAN. Transport solutions adopting the C-RAN architecture require infrastructure connectivity within the RAN known as FH. Although C-RAN overcomes traditional RAN limitations, it requires the support of new operational network services over the transport network to meet the challenges of emerging services. This can be addressed decomposing traditionally monolithic RAN processing functions stack to a set of different units referred to as RAN split options that can relax the corresponding transport network requirements regarding overall capacity, delay and synchronisation. The optimal split option depends on parameters such as supported services, service requirements, technology and protocols of the FH and BH, etc. To maximize coordination and resource sharing gains, it is proposed to support BH and FH jointly in a common infrastructure. In view of this several solutions have been proposed including both wired and wireless approaches. In terms of wired transport network solutions, a variety of technical approaches have been proposed including Programmable Elastic Frame based optical networks, Programmable Metro Network exploiting disaggregated Edge Nodes, Space Division Multiplexing and Ethernet transport. In a complementary manner, alternative wireless transport solutions are also proposed including mmWave approaches exploiting the concept of multi-tenant small cells with integrated access and BH, satellite BH as well as fibre wireless point-to-multipoint solutions. Finally, specific examples of data plane programmability of the data path in non-RAN segments such as the Edge Network, the Transport Network and the CN are presented. The purpose for this is to enable network traffic/slice Quality of Service (QoS) control in the data plane, and thus enable QoS-aware network slicing. These include Stateful Packet Processing in Hardware and Segment Routing.

The analysis of the architecture of multiple management and orchestration systems in Chapter 5 revealed that there is considerable consensus on how to structure such systems at a high level of abstraction. The architecture defined by ETSI NFV is generally adopted and provides a stable base to MANO systems. There is, however, considerable freedom in how to realize such a management and orchestration system. This freedom is partially driven by opportunity (evolving standards, but also technological opportunities like the growing importance of accelerator platforms and containers), partially by an ongoing discussion for a best possible solution (for example, the options of flat, hierarchical, or recursive orchestration, or the approach towards DevOps processes). Evaluation of different MANO approaches is now starting, and it will benefit from community-wide agreement on typical work load assumptions a MANO system will have to deal with (e.g., how many services will have to be instantiated in a typical operator network per second, per minute, or per day). Hence, there is still an opportunity for additional work in the MANO area, both on the architectural level as well as from a perspective of having to base the work on concrete, quantifiable scenarios and measurements.

Chapter 6 introduces the approach to evaluate how the architecture presented in this paper has been deployed in the context of the different use cases in the projects. To achieve this the project have answered a number of questions with respect to the main characteristics of a 5G system such as the use of 5G new radio, the frequencies used, the deployment of a MEC, the type of transport network technology used, the use of 5G core and generally the use of automated or autonomic methods for the management and orchestration of the deployed network and services.

Furthermore it provides an introduction to the evaluation of the programme performance KPIs. In the context of KPI validation it presents an approach to measure the performance KPIs at the borders of the main network architecture segments. Further analysis is needed to quantify the performance of the systems, which is largely part of ongoing work in the Test, Measurement and KPI Validation work group. In addition further evaluation by vertical actors is needed to qualify whether the deployed systems have indeed met the vertical actor's requirements.

Finally, in Chapter 7, standards contributions from 5G PPP Phase 2 projects were analyzed. It was noted that the contributions were evenly distributed across the four main architectural areas considered in this report in Chapters 2 to 5, with 3GPP and ETSI being the SDOs mostly targeted by the projects. However, it was also noted that there are efforts to promote the role of vertical industries in further enhancing the standards impacts for 5G PPP Phase 3 projects.

9 Glossary

This section defines the key terminology used in this White Paper. The definitions of terms are structured according to their area, such as virtualization related or business related. The terms defined here are the most relevant ones, especially those that have different definitions by various standardization developing organizations.

9.1 General Terms

Information model (IM): An abstraction and representation of the entities in a managed environment, their properties, attributes and operations, and the way they relate to each other. It is independent of any specific repository, software usage, protocol, or platform. [9-57]

Data model: A mapping of the contents of an information model into a form that is specific to a particular type of data store or repository [9-57]. A "data model" is basically the rendering of an information model according to a specific set of mechanisms for representing, organizing, storing and handling data. It has three parts:

- A collection of data structures such as lists, tables, relations, etc.
- A collection of operations that can be applied to the structures such as retrieval, update, summation, etc.
- A collection of integrity rules that define the legal states (set of values) or changes of state (operations on values).

Policy: Policy [9-1] can be defined from two perspectives:

- A definite goal, course or method of action to guide and determine present and future decisions. "Policies" are implemented or executed within a particular context (such as policies defined within a business unit).
- Policies as a set of rules to administer, manage, and control resources (of networking, computational and storage), services and their lifecycles, software as well as hardware of the system.

NOTE: These two views are not contradictory since individual rules may be defined in support of business goals [9-57].

Service: The behaviour or functionality provided by a network, network element or host. To completely specify a "service", one must define the "functions to be performed ..., the information required ... to perform these functions, and the information made available by the element to other elements of the system". Policy can be used to configure a "service" in a network or on a network element/host, invoke its functionality, and/or coordinate services in an inter-domain or end-to-end environment. [9-57]

9.2 Network function virtualization related

The central concepts around network function virtualization and network services are based on the definitions of ETSI NFV.

Network Function (NF): functional block within a network infrastructure that has well-defined external interfaces and well-defined functional behaviour. [9-15]

Network Service (NFV-NS): composition of Network Functions and defined by its functional and behavioural specification. [9-15]

NOTE: “The Network Service contributes to the behaviour of the higher layer service, which is characterized by at least performance, dependability, and security specifications. The end-to-end network service behaviour is the result of the combination of the individual network function behaviours as well as the behaviours of the network infrastructure composition mechanism.” [9-15]

NOTE: A network service can be seen as a set of VNFs or PNFs, connected by VLs as defined in a VNFFG.

Network Service Descriptor (NSD): template that describes the deployment of a Network Service including service topology (constituent VNFs and the relationships between them, Virtual Links, VNF Forwarding Graphs) as well as Network Service characteristics such as SLAs and any other artefacts necessary for the Network Service on-boarding and lifecycle management of its instances. [9-15]

NOTE: The NSD includes a number of deployment flavors, each referencing deployment flavors of all or a subset of the NFV-NS’s constituent VNFs and Virtual Links. The NSD also provides a list of pointers to the descriptors of its constituent VNFs (i.e. VNFDs) and additional information on the connectivity between them together with the traffic forwarding rules.

Network Service Instance (NFV-NSI): refers to an instance of a network service (NFV-NS).

NFVI as a Service (NFVIaaS): The tenant is offered a virtual infrastructure including associated resources (networking/computing/storage) under its full control in which it can deploy and manage its own NFV network services on top of it. It is assumed that the tenant will deploy its own MANO stack. This is probably the most usual service consumed by M(V)NOs, given that they have the knowledge and need to customize their communication service offering to their own customers. Resources could be virtual cores, storage, virtual nodes and links, etc.

NOTE: The tenant can deploy and connect VMs on these resources under its own control.

NOTE: NFVIaaS includes the provision of network slices or network slice subnets as a service.

Network Service as a Service (NSaaS): Provide to a tenant the possibility to define and instantiate a network service.

NF forwarding graph (NF FG): graph of logical links connecting NF nodes for the purpose of describing traffic flow between these network functions. [9-15]

Physical Application (PA): implementation of a VA via a tightly coupled software and hardware system.

NOTE: analogous to PNF.

NOTE: may include devices such as cameras, smart city sensors, etc.

Physical Network Function (PNF): implementation of a NF via a tightly coupled software and hardware system. [9-15]

VA Forwarding Graph (VA FG): Forwarding graph among VA, VNF, PA, PNF nodes.

Virtual Application (VA): more general term for a piece of software which can be loaded into a Virtual Machine. [9-15]

Virtual link (VL): set of connection points along with the connectivity relationship between them and any associated target performance metrics (e.g. bandwidth, latency, QoS). [9-15]

Virtualised Network Function (VNF): implementation of an NF that can be deployed on a Network Function Virtualisation Infrastructure. [9-15]

Virtualised Network Function Component (VNFC): internal component of a VNF providing a defined sub-set of that VNF's functionality, with the main characteristic that a single instance of this component maps 1:1 against a single Virtualisation Container. [9-15]

Virtualised Network Function Descriptor (VNFD): configuration template that describes a VNF in terms of its deployment and operational behaviour, and is used in the process of VNF onboarding and managing the lifecycle of a VNF instance. [9-15]

VNF Forwarding Graph (VNF FG): NF forwarding graph where at least one node is a VNF. [9-15]

9.3 Network slice related

Network slice (NS): A network slice is a complete logical network over a shared compute, storage and network infrastructure. E.g. a network operator can build a network slice including an Access Network (AN) and a Core Network (CN) to enable communication services.

Network slice instance (NSI): a set of network functions and the resources for these network functions which are arranged and configured, forming a complete logical network to meet certain network characteristics [9-6].

NOTE: There are different ways of deploying network slices. In the context of ETSI NFV, a network slice instance could be deployed as a NFV Network Service instance (NFV-NSI). In this context, different slices can be deployed as instances of the same type of NFV-NS with different deployment flavours or instances of different types of NFV-NS. In an NFV framework, creating a network slice will typically involve filling an NSD and requesting the NFV Orchestrator to instantiate an NFV-NS according to the contents of its NSD and selected deployment flavour.

Network slice subnet instance (NSSI): a set of network functions and the resources for these network functions which are arranged and configured to form a logical network (sub-network) [9-6].

NOTE:

- A NSI may include one or more NSSIs, which can include one or more VNFs or PNFs.
- A NSSI can be shared by multiple NSIs. In this case, the shared NSSIs have to be configured appropriately to provide proper isolation and separation.

9.4 Vertical service related

Vertical: the stakeholder belonging to an industrial sector and consuming services (defined in Section 9.6). MVNOs are considered a special type of vertical.

NOTE: The existence of network slices is transparent to the vertical and it is fully under the control of the Service Provider how to handle them, including, for instance, mapping services into network slices.

Vertical Service (VS): From a business perspective, it is a service focused on a specific industry or group of customers with specialized needs (e.g., automotive services, entertainment services, e-health services, industry 4.0).

From a technical point of view, it is a composition of general functions as well as network functions

Vertical Service Blueprint (VSB): A parameterized version of a Vertical Service Descriptor, where parameters have to be provided to provide a complete VSD, which is ready to be instantiated.

NOTE: There can be a wide range of parameters. The parameters can be used to express requirements of the vertical service, but also management related parameters such as file locations of virtual machine images or the priority of a service. A subset of parameters to express requirements are: Bitrate of VAs and the connecting links, round-trip time among two VAs, geographical area to be covered by the vertical service.

Vertical Service Descriptor (VSD): A description of the deployment of a vertical service including service topology (constituent VAs and the relationships between them, Virtual Links, VNF Forwarding Graphs) as well as vertical service characteristics such as SLAs and any other artefacts necessary for the vertical service on-boarding and lifecycle management of its instances.

NOTE: A VSD may still contain instance-specific parameters to be provided at instantiation time. This is similar to parameters provided at instantiation time of VNFs.

9.5 Multi-access edge computing related

The central concepts around multi-access edge computing are based on the definitions of ETSI MEC [9-28] and recent draft integrating NFV and MEC [9-31]. Following the renaming of mobile edge computing to multi-access edge computing, the definitions from [9-28] have been changed accordingly.

Multi-access edge application (MEA): application that can be instantiated on a multi-access edge host within the multi-access edge system and can potentially provide or consume multi-access edge services. [9-28]

Multiple-access Edge Application Orchestrator (MEAO): It has the same functions as MEO, excepting that it should use the NFVO to instantiate the virtual resources for the MEA as well as for the MEP.

Multiple-access Edge Host (MEC Host): It provides the virtualization environment to run MEC applications, while it interacts with the mobile network entities, via the MEP platform, to provide MES and offload data to MEA.

Multiple-access Edge Orchestrator (MEO): The MEO is in charge of the orchestration and the instantiation of MEA.

Multiple-access Edge Platform Manager (MEPM): It is in charge of the life-cycle management of the deployed MEA. The MEPM is in charge of the MEP configuration, such as the MEC application authorization, the traffic type need to be offloaded to the MEC application, DNS redirection, etc.

Multiple-access Edge Platform Manager – NFV (MEPM-V): The virtualized version of the MEPM delegates the LCM of MEA to one or more VNFMs, and keeps the MEP configuration.

Multi-access edge platform (MEP): collection of functionality that is required to run multi-access edge applications on a specific multi-access edge host virtualisation infrastructure and to enable them to provide and consume multi-access edge services, and that can provide itself a number of multi-access edge services. [9-28]

Multi-access edge service (MES): service provided via the multi-access edge platform either by the multi-access edge platform itself or by a multi-access edge application. Examples of MES provided by the MEP are the radio network information service or the location service. [9-28]

In terms of orchestration, some of the MEC concepts have an analogy in NFV, e.g. the MEAO and MEO taking a similar role as the NFVO in orchestrating virtual functions. In MEC these functions are the MEAs, whereas in NFV these are the VNFs. In addition to these similar concepts, MEC provides also predefined services for applications related to mobile devices.

9.6 Business logic/stakeholder related

Services: Service may be provided by different types of service providers, e.g., Communication Service Provider offering traditional telecom services, Digital Service Provider offering digital services such as enhanced mobile broadband and IoT to various vertical industries, or Network Slice as a Service (NSaaS) Provider offering a network slice as a service to its customers. Also, the services offered to verticals may differ to which degree they can be managed by the vertical itself.

Managed Vertical Service (MVS): Vertical services that are fully deployed and managed by the SP and consumed as such by the vertical.

Unmanaged Vertical Service (UVS): Vertical services that are deployed by the SP (i.e., instantiating VNFs and their connectivity), i.e. the lifecycle of the service is done by the SP. But the service logic is only partially or not at all managed by the SP, instead it is managed by the vertical. This includes the configuration of VNF internals to control the logic of the vertical services at service level.

NOTE: If also the lifecycle is managed by the vertical, the service follows the NFVIaaS model as defined before.

Service Customer (SC): uses services that are offered by a Service Provider (SP). In the context of 5G, vertical industries are considered as one of the major SCs.

Service Provider (SP): comprises three sub-roles, depending on the service offered to the SC: traditional Communication Service Provider, Digital Service Provider, or Network Slice as a Service (NSaaS) Provider. SPs design, build and operate services using aggregated network services.

Network Operator (NOP): in charge of orchestrating resources, potentially from multiple virtualised infrastructure providers (VISP). The NOP uses aggregated virtualised infrastructure services to design, build, and operate network services that are offered to SPs.

Virtualisation Infrastructure Service Provider (VISP): Provides virtualised infrastructure services and designs, builds, and operates virtualisation infrastructure(s) (3GPP28.801, 2017). The infrastructure comprises networking (e.g., for mobile transport) and computing resources (e.g., from computing platforms).

Data Centre Service Provider (DCSP): Provides data centre services and designs, builds and operates its data centres. A DCSP differs from a VISP by offering “raw” resources (i.e., host servers) in rather centralised locations and simple services for consumption of these raw resources. A VISP rather offers access to a variety of resources by aggregating multiple technology domains and making them accessible through a single API.

9.7 DevOps-related terms and roles

This section will detail some of the high-level development-oriented terms from above. These terms here support a more fine-grained perspective on where and in which **roles** software artefacts (of whatever form, for example, but not limited to, applications, general functions or services such as network functions or network services) are developed and deployed. The notion of a role of

course entails that these roles can be mapped in different ways to commercial entities; the same company can and often does assume multiple roles (e.g., often function, service, and infrastructure developer will overlap), sometimes even different subsets of roles in different business relationships. There is, however, little mandate that some of these roles need to coexist in the same company (historic evidence or traditional setups of business models notwithstanding); rather, there is some plausibility to mandate that some roles are kept separated and potentially realized by a neutral, trusted third party (e.g., validation entities). But of course, these are at best recommendations; the concrete development should be left over to market developments and not be imposed by a technical document.

Note that with cloud-computing-based developments like Infrastructure-as-Code, the differentiation between software and infrastructure becomes less and less relevant.

End user: The actual user and ultimate beneficiary of an application or a service. It can be a private person or a technical device operating on behalf of but independent from an actual person.

Function developer: A developer of functions. A function here is an executable entity in an atomic sense of NFV VNFs, of MEC, or even in the sense of application-oriented microservices. The executable form can be provided in different ways (e.g., source code, virtual machine image, container description, executable process, JAR file, etc.); there is no mandate made here. A function developer will typically make some functional claims about such a function.

Service developer: A service developer uses existing functions and existing services to develop the description of new services. These descriptions are made available as network service descriptors (NSDs) in one or several formats. Similar to a function, a service description can comprise some functional claims. The distinction between function developer and service developer is fluid and cannot always clearly be made.

Application developer: Given that the difference between “service” and “application” is more or less disappearing, service developers and application developers can be regarded as the same role.

Infrastructure developer: This role comes in two flavours: the physical infrastructure developer and the virtual infrastructure developer. The physical infrastructure developer is concerned with building up hardware IT infrastructure, such as data centres. With the trend towards infrastructure-as-code, the virtual infrastructure on top of which an application or a service is intended to execute is becoming a textual artefact. This artefact can describe aspects like desired network topology (e.g., setup of Layer2 networks), required resources like storage, required virtualizers, etc. Since (virtual) infrastructure for a service and service itself are tightly related, often, the roles of service developer and infrastructure developer will overlap, but they are indeed separate roles and aspects. Also, it is worthwhile to point out that such infrastructure descriptions exist already in today’s description formats but are usually lumped in together with the function/service description itself; even though structurally, they really form separate artefacts.

Functional claim: Statements about the intended function of a function or a service, claims about its correct operation. These claims are functional in the sense that they only pertain to which functions are executed/which service is provided, they are *not* claims about how well they do that (see non-functional claims).

Non-functional claims: Non-functional claims are statements about how well a function/service performs, typically with respect to some quantitative metric (achieved throughput, delay, error rates, availability, reliability, ...) or with respect to some non-functional yet non-quantitative properties like security (ability to withstand attacks) or maintainability (ability to be upgraded, possibly while in operation). In a loose sense, all these metrics can be regarded as realizations of LaPrie’s dependability concept [9-63].

NOTE: Some of these claims – in particular, quantitative performance claims – make no sense about a function in isolation but only when considered together with a particular

infrastructure on top of which a function/a service is assumed to be executed. Hence, such a non-functional claim pertains always to a tuple of (function/service description; infrastructure description) – a prime reason why the notion of an infrastructure description is important.

Non-functional claims developer: while it makes sense to assume that functional claims are made directly by the function/service developer, non-functional claims are more complex and can only be developed together with an assumption about the used infrastructure. Hence, it stands to reason to identify a separate role of non-functional claims development. Obviously, this role is likely to overlap with other developer roles in practice.

Validation and verification entity: With claims about functional and non-functional properties of a function/service in place, an operator could, in principle, decide how many resources are needed for a service to achieve which service level, typically also assuming something about the load, number of users, etc. However, it is not clear whether these claims are trustworthy (they typically are not, even if the function software & claims were developed inhouse by the operator itself). Hence, an additional role is needed to validate and verify such claims. The methods to do so are manifold (e.g., simulation, experiments in testing infrastructure, even formal proofs, etc.) and are outside the scope of a whitepaper. In fact, there are additional artefacts necessary in detail (e.g., a description which type of validation infrastructure are available in this role). Also, this role is typical example to be realized by a neutral third party (neutral between developers and operators) and working on behalf of either of these parties, providing “validated claims”.

Tenant: A tenant owns and typically also orders a service or an application. There is no notion of a tenant belonging to any particular industry (be it vertical or otherwise), is a commercial entity (a private person can of course act as a tenant), or whether the tenant intends to make the service/application publically available or use it for its own private or internal purposes.

Physical infrastructure provider: An infrastructure provider operates actual, physical infrastructure. It uses infrastructure descriptions to create virtual infrastructure on top of which functions/services/applications can be deployed and used. Infrastructure can comprise networking, computational, storage infrastructure, or any subset of these. An infrastructure provider can use arbitrary tools to ensure that the desired virtual infrastructure has the desired properties. For example, it can use SDN techniques to provide a desired topology of servers and access points, making a virtual network available to its customers.

Infrastructure provider: Generalizes the physical infrastructure provider by allowing to offer either physical resources or virtual infrastructure, turn them into a desired virtual infrastructure and sell it. Effectively makes the notion of an infrastructure provider recursive by allowing to resell it.

Market: A structure where functions, service descriptions, actually running services, infrastructure descriptions, (real or virtual) infrastructures and other artefacts can be traded, under various contracts. Markets can be simplistic (just lists of items with contact points) or very sophisticated (with search facilities, creating desired services out of components, etc.).

Market provider: An entity realizing a market.

Operator: An operator is a business model rather than a role. It is nevertheless characterized by typical combinations of roles, in particular a combination of infrastructure provider and service provider plus customer-relationship manager and billing agency. Other combinations of roles are possible and some of them can also emerge with the proliferation of 5G networks.

9.8 Specific terms

Abstracted Resource/ Resource abstraction: Limited description of a resource with intention to hide certain parameters (such as quantity, vendors, location of the resource, etc.) and secure enough to be shared with other administrative domains.

Abstracted Service/ Service abstraction: Limited description of a service with intention to hide certain parameters (such as used resources, virtual links, interconnections etc.) and secure enough to be shared with other administrative domains.

Administrative domain: is a collection of resources and/or services owned and operated by a single administrative organization.

Resources Federation: Set of resources can be offered by a provider domain under pre-agreed terms and conditions; available resources potentially to be used by a consumer Service Provider domain, with certain pre-agreed terms and conditions. In this case, the resources is owned by the provider domain but managed by the consumer Service Provider, similar to NFVI as a Service case in the context of ETSI NFV.

Services Federation: Set of services can be offered by a provider domain to other potential consumer domains, under pre-agreed terms and conditions. Different to the resource federation case, the provider domain is fully in charge of the services along with their life cycle management and required resources for deploying them within the provider domain and offer them to the consumer domains, similar to Network Service as a Service case in the context of ETSI NFV.

Consumer domain: Administrative domain that demands resources or services from other administrative domains. Note that a consumer domain can use these consumed resources or services and provide them again in another context, acting as a provider domain again.

Federated Resources: Resources is fully controlled and managed (i.e., instantiation, reservation, allocation, scaling up/down and release) by a consumer domain, but owned by a provider domain (operator or infrastructure provider). The consumer domain is allowed (by the provider domain) to manage and use the resources based on pre-agreed terms and conditions (SLAs). In this case, the consumer SP uses NFV (abstracted) virtual resources offered by the peer SP. This may be the case when an end-to-end NFVIaaS service is built by combining virtual resources belonging to multiple SP administrative domains.

Federated Services: Services managed by a consumer domain, but owned by a provider domain. The consumer domain is allowed (by the provider domain) to manage and use the services based on pre-agreed terms and conditions (SLAs). In this case, the consumer SP uses NFV network services offered by the peer SP. This may be the case when an end-to-end service is split into constituent services that are deployed in multiple SP administrative domains.

Federation is a mechanism for integrating resources and services from multiple administrative domains at different granularity according to their pre-agreed administrative relations and signed contract agreement of sharing relevant information between them [9-58].

Local Repository: Database (in an administrative domain) that holds information for available resources for federation, catalogue of services/abstracted services, provided by other provider domains.

Provider domain: Administrative domain that offers resources or services to other administrative domains. Note that the provided resources or services can be based on resources or services consumed from some other domain as well.

Service catalogue: Composed set of services and/or service abstractions offered by a provider domain to other potential consumer domains using mutual taxonomy and agreed usage terms

(SLAs). In case of federation, the composed service catalogue is shared and continuously updated between the federated administrative domains.

Technology domain: is a collection of resources that are part of a single technology (system) and belong to a single administrative domain. The internal structure is defined and operated according to the technology definitions and standards. One or more technology domains can be part of an administrative domain.

9.9 References

- [9-1] 5G-TRANSFORMER, D1.1, Report on vertical requirements and use cases, November 2017.
- [9-2] 5G-TRANSFORMER, D2.1, Definition of the Mobile Transport and Computing Platform, March 2018.
- [9-3] 5G-TRANSFORMER, D3.1, Definition of vertical service descriptors and SO NBI, March 2018.
- [9-4] 5G-TRANSFORMER, D4.1, Definition of service orchestration and federation algorithms, service monitoring algorithms, March 2018.
- [9-5] 5G-TRANSFORMER, D5.1, Definition of vertical testbeds and initial integration plans, May 2018.
- [9-6] 3GPP TR 28.801, V15.0.0, Telecommunication management; Study on management and orchestration of network slicing for next generation network, 2017.
- [9-7] 3GPP TS 23.501, V15.0.0, System Architecture for the 5G System; Stage 2, 2017.
- [9-8] 3GPP TR 23.714, V14.0.0, Study on control and user plane separation of EPC nodes, 2016.
- [9-9] 3GPP TS28.530, V0.4.0, Telecommunication management; Management of 5G networks and network slicing; Concepts, use cases and requirements, 2017.
- [9-10] NHTSA, “Preliminary regulatory impact analysis - FMVSS No. 150 Vehicle-to-Vehicle Communication Technology for Light Vehicles,” December 2016.
- [9-11] G. Avino, M. Malinverno, C. Casetti, C. F. Chiasserini, F. Malandrino, M. Rapelli, G. Zennaro “Support of Safety Services through Vehicular Communications: The Intersection Collision Avoidance Use Case”.
- [9-12] B. Chatras, U. S. Tsang Kwong and N. Bihannic, "NFV enabling network slicing for 5G," 2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN), Paris, 2017, pp. 219-225.
- [9-13] TeleManagement Forum. “TeleManagement Forum Information Framework (SID): GB922_Addendum_4SO_Service_Overview_R9-5_v9-7”.
- [9-14] ETSI GS NFV-MAN 001, “Network Functions Virtualisation (NFV); Management and Orchestration”, v1.1.1, 2014.
- [9-15] ETSI GS NFV 003, “Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV”, v1.2.1, Dec. 2014.
- [9-16] ETSI GS NFV-IFA 005, “Network Function Virtualisation (NFV); Management and Orchestration; Or-Vi reference point – Interface and Information Model Specification”, v2.1.1, 2016.
- [9-17] ETSI GS NFV-IFA 006, “Network Function Virtualisation (NFV); Management and Orchestration; Vi-Vnfm reference point – Interface and Information Model Specification”, v2.1.1, 2016.
- [9-18] ETSI GS NFV-IFA 007, “Network Function Virtualisation (NFV); Management and Orchestration; Or-Vnfm reference point – Interface and Information Model Specification” v2.1.1, 2016.
- [9-19] ETSI GS NFV-IFA 008, “Network Functions Virtualisation (NFV); Management and Orchestration; Ve-Vnfm reference point - Interface and Information Model Specification”, v2.1.1, 2016.
- [9-20] ETSI GS NFV-IFA 010, Management and Orchestration; Functional requirements specification”, v2.4.1, 2018.
- [9-21] ETSI GS NFV-IFA 011, “Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; VNF Packaging Specification” v2.3.1, August 2017.
- [9-22] ETSI GS NFV-IFA 012, draft, Management and Orchestration Os-Ma-Nfvo reference point – Application and Service Management Interface and Information Model Specification, v0.11.0, 2017.

- [9-23] ETSI GS NFV-IFA 013, “Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Os-Ma-Nfvo reference point – Interface and Information Model Specification” v2.3.1, August 2017.
- [9-24] ETSI GS NFV-IFA 014, V2.4.1, Management and Orchestration; Network Service Templates Specification, 2018.
- [9-25] ETSI GR NFV-IFA 022, Management and Orchestration; Report on Management and Connectivity for Multi-Site Services”, v0.8.2, 2018.
- [9-26] ETSI GS NFV-IFA 028, Management and Orchestration; Report on architecture options to support multiple administrative domains”, v3.1.1, 2018.
- [9-27] ETSI GR NFV-EVE 012 V3.1.1, Evolution and Ecosystem; Report on Network Slicing Support with ETSI NFV Architecture Framework, 2017.
- [9-28] ETSI GS MEC 001 V1.1.1, Mobile Edge Computing (MEC): Terminology, March 2016.
- [9-29] ETSI GS MEC 003, “Mobile Edge Computing (MEC); Framework and Reference Architecture”, v1.1.1, March 2016.
- [9-30] ETSI GS MEC 010-2, “Mobile Edge Computing (MEC); Mobile Edge Management; Part 2: Application lifecycle, rules and requirements management”, v1.1.1, July 2017.
- [9-31] ETSI GR MEC 017, “Mobile Edge Computing (MEC); Deployment of Mobile Edge Computing in an NFV Environment”, v1.1.1, February 2018.
- [9-32] Antonio de la Oliva et al., “5G-TRANSFORMER: Slicing and Orchestrating Transport Networks for Industry Verticals”, accepted by IEEE Communications Magazine, 2018.
- [9-33] C. Casetti et al., Network Slices for Vertical Industries, 1st Workshop on Control and management of Vertical slicing including the Edge and Fog Systems (COMPASS), Barcelona, IEEE, 2018.
- [9-34] Xi Li et al., “Service Orchestration and Federation for Verticals”, 1st Workshop on Control and management of Vertical slicing including the Edge and Fog Systems (COMPASS), Barcelona, IEEE, 2018.
- [9-35] P. Iovanna et al., “5G Mobile Transport and Computing Platform for verticals”, 1st Workshop on Control and management of Vertical slicing including the Edge and Fog Systems (COMPASS), Barcelona, IEEE, 2018.
- [9-36] Horizon 2020, 5GPPP: 5G-Crosshaul project, Available at: <http://5g-crosshaul.eu/>
- [9-37] D. King and A. Farrell, “A PCE-Based Architecture for Application-Based Network Operations”, IETF RFC 7491, March 2015.
- [9-38] M. Björklund, Martin, “YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)”, IETF RFC6020, October 2010.
- [9-39] A. Bierman et al, “Restconf Protocol”, IETF RFC 8040, January 2017.
- [9-40] L. Qiang et al., “Technology Independent Information Model for Network Slicing”, draft-qiang-coms-netslicing-information-model-02, Work in progress, Expires July 30, 2018.
- [9-41] RFC 2119, “Key words for use in RFCs to Indicate Requirement Levels”, IETF, 1997.
- [9-42] Topology and Orchestration Specification for Cloud Applications Version 1.0. 25 November 2013. OASIS Standard. [Online] <http://docs.oasis-open.org/tosca/TOSCA/v1.0/os/TOSCA-v1.0-os.html>
- [9-43] TOSCA Simple Profile for Network Functions Virtualization (NFV) Version 1.0, Edited by Shitao Li and John Crandall. 11 May 2017. OASIS Committee Specification Draft 04, [Online] <http://docs.oasis-open.org/tosca/tosca-nfv/v1.0/tosca-nfv-v1.0.pdf>
- [9-44] The TOSCA Cloud Service Archive (CSAR), [Online] <https://www.oasis-open.org/committees/download.php/46057/CSAR%20V0-1.docx>
- [9-45] YAML, [Online], <http://yaml.org/>
- [9-46] SONATA NFV, <http://www.sonata-nfv.eu/>
- [9-47] SONATA D2.2. Architecture Design, [Online] <http://www.sonata-nfv.eu/content/d22-architecture-design-0>
- [9-48] Description of Network Slicing Concept by NGMN Alliance, NGMN 5G P1 Requirements & Architecture Work Stream End-to-End Architecture, version 1.0, January 2016.

- [9-49] NGMN Alliance, “Description of Network Slicing Concept”, v.1.0.8, September 2016.
- [9-50] NGMN Alliance, NGMN 5G White Paper, February 2015.
- [9-51] Open Source MANO. <https://osm.etsi.org/>
- [9-52] J. Baranda et al., “Resource Management in a Hierarchically Controlled Multidomain Wireless/Optical Integrated Fronthaul and Backhaul Network”, IEEE NFV-SDN conference, November 2017.
- [9-53] 5GEx, <https://5G PPP.eu/5gex/>
- [9-54] 3GPP TS 29.060 V15.1.0 (2017-12), General Packet Radio Service (GPRS); GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface.
- [9-55] IETF, RFC7348, Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks, August 2014.
- [9-56] Multi-access Edge Computing, <http://www.etsi.org/technologies-clusters/technologies/multi-access-edge-computing>
- [9-57] IETF RFC3198, Terminology for Policy-Based Management, November 2001.
- [9-58] L. Cominardi, "Multi-domain federation: scope, challenges, and opportunities," Workshop in 3rd IEEE Conference on Network Softwarization, Bologna, Italy, July 2017.
- [9-59] C. J. Bernardos, A. De La Oliva, F. Giust, J.C. Zuniga, A. Mourad, “Proxy Mobile IPv6 extensions for Distributed Mobility Management”, September 2018, IETF draft (work-in-progress), draft-bernardos-dmm-pmipv6-dlif-01.
- [9-60] C. J. Bernardos, et al., “Multi-domain Network Virtualization”, September 2018, IETF draft (work-in-progress), draft-bernardos-nfvrg-multidomain-04.
- [9-61] L. Geng, et al., “COMS Architecture”, September 2018, IETF draft (work-in-progress), draft-geng-coms-architecture-02.
- [9-62] L. Geng, et al., “Problem Statement of Common Operation and Management of Network Slicing”, September 2018, IETF draft (work-in-progress), draft-geng-coms-problem-statement-03.
- [9-63] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, “Basic concepts and taxonomy of dependable and secure computing”, IEEE Transactions on Dependable and Secure Computing, Jan-March 2004.

10 List of Contributors

Name	Company / Institute / University	Country
Editorial Team		
<i>Overall Editors</i>		
Simone Redana	Nokia Bell Labs 5G PPP Architecture WG Chairman	Germany
Ömer Bulakci	Huawei Technologies Duesseldorf GmbH 5G PPP Architecture WG Vice-Chairman	Germany
<i>Section 2 Editor</i>		
Christian Mannweiler	Nokia Bell Labs	Germany
<i>Section 3 Editors</i>		
Laurent Gallo	Orange	France
Apostolos Kousaridas	Huawei Technologies Duesseldorf GmbH	Germany
<i>Section 4 Editors</i>		
David Navrátil	Nokia Bell Labs	Finland
Anna Tzanakaki	University of Bristol	UK
Jesús Gutiérrez	IHP	Germany
<i>Section 5 Editors</i>		
Holger Karl	Paderborn University	Germany
Peer Hasselmeyer	NEC Laboratories Europe GmbH	Germany
<i>Section 6 Editor</i>		
Anastasius Gavras	Eurescom GmbH	Germany
<i>Section 7 Editors</i>		
Stephanie Parker	TrustIT	Italy
Edward Mutafungwa	Aalto University	Finland
Contributors		

Anastasios Zafeiropoulos	UBITECH	Greece
Anastasius Gavras	Eurescom GmbH	Germany
Anna Tzanakaki	University of Bristol	UK
Antonino Albanese	Italtel S.p.A.	Italy
Apostolos Kousaridas	Huawei Technologies Duesseldorf GmbH	Germany
Avi Weit	IBM	Israel
Bessem Sayadi	Nokia Bell Labs	France
Boris Tiomela Jou	Airbus	France
Carlos J Bernardos	University of Carlos III Madrid	Spain
Chafika Benzaid	Aalto University	Finland
Christian Mannweiler	Nokia Bell Labs	Germany
Dan Warren	Samsung Research UK	UK
Daniel Camps-Mur	I2CAT Foundation	Spain
David Breitgand	IBM	Israel
David Gutierrez Estevez	Samsung	UK
David Navratil	Nokia Bell Labs	Finland
De Mi	University of Surrey	UK
Diego Lopez	Telefonica	Spain
Dimitrios Klonidis	Athens Information Technology	Greece
Edward Mutafungwa	Aalto University	Finland
Eleni Fotopoulou	UBITECH	Greece
Emmanouil Kafetzakis	ORION INNOVATIONS P.C.	Greece
Emmanouil Pateromichelakis	Huawei Technologies Duesseldorf GmbH	Germany
Erez Biton	Nokia	Israel
Fasil B. Tesema	Nomor Research GmbH	Germany

George Kalfas	Aristotle University of Thessaloniki	Greece
Holger Karl	Paderborn University	Germany
Jens Bartelt	Airrays GmbH	Germany
Jesús Gutiérrez	IHP	Germany
John Cosmas	Brunel University	UK
John Thomson	Onapp	UK
Jordi Joan Giménez	Institut für Rundfunktechnik GmbH	Germany
Jose M. Alcaraz Calero	University of the West of Scotland	UK
Josep Mangués-Bafalluy	Centre Tecnològic de Telecomunicacions de Catalunya	Spain
Kostas Katsalis	Huawei Technologies Duesseldorf GmbH	Germany
Laurent Gallo	Orange	France
Marco Gramaglia	University of Carlos III Madrid	Spain
Maria Rita Spada	Wind TRE S.p.A.	Italy
Mukhald Salih	Brunel University	UK
Navid Nikaein	EURECOM	France
Nawar Jawad	Brunel University	UK
Nebojsa Maletic	IHP	Germany
Nicolas Chuberre	Thales Alenia Space	France
Ömer Bulakci	Huawei Technologies Duesseldorf GmbH	Germany
Panagiotis Demestichas	Incelligent	Greece
Peer Hasselmeyer	NEC Laboratories Europe GmbH	Germany
Qi Wang	University of the West of Scotland	UK
Qing Wei	Huawei Technologies Duesseldorf GmbH	Germany
Refik Fatih Ustok	NETAŞ	Turkey
Rolf Blom	RISE SICS	Sweden
Salvatore Pontarelli	CNIT	Italy

Selçuk Keskin	NETAŞ	Turkey
Stefano Salsano	CNIT/Universita' di Roma Tor Vergata	Italy
Stephanie Parker	TrustIT	Italy
Thomas Deiss	Nokia	Germany
Ugur Acar	NETAŞ	Turkey
Xi Li	NEC Laboratories Europe GmbH	Germany
Yue Zhang	University of Leicester	UK

Contributing 5GPPP Projects

Phase 2 Projects

<https://5g-ppp.eu/5g-ppp-phase-2-projects/>



Phase 3 Projects

<https://5g-ppp.eu/5g-ppp-phase-3-projects/>



11 List of 5G PPP Phase 2 Contributions to Standards

Project Name	Targeted SDO	Title or short description of contribution related to overall architecture	Status
5G CAR	3GPP-SA6	New key issue on interaction between V2X application and 3GPP system for V2X application and QoS adaptation	S6-180782
5G CAR	3GPP-SA6	Solution proposal for key issue #13 communicating application requirements from the V2X application server	S6-181048
5G CAR	3GPP-SA6	Procedures for service negotiation	S6-181352
5G MoNArch	3GPP-SA2	New Key Issue: Network slicing for eV2X	S2-180147
5G MoNArch	3GPP-SA2	23.726: FS_ETSUN (Enhancing Topology of SMF and UPF)/23.726 Scope	S2-181046
5G MoNArch	3GPP-SA2	NWDA-assisting E2E QoS Assurance	S2-183634
5G MoNArch	3GPP-SA2	Solution for AF Data Exposure to/from NWDAF	S2-183637
5G MoNArch	3GPP-SA2	Use case on UE-driven analytics sharing	S2-185816 (S2-185290)
5G MoNArch	3GPP-SA2	Key issue for NWDA-assisted determination of areas with oscillation of network conditions	S2-185277
5G MoNArch	3GPP-SA2	Solution for Key Issue 4: interactions with OAM for network data collection	S2-185279
5G MoNArch	3GPP-SA2	Updates on text of Key Issue 1 and general architectural assumptions with exposure to OAM	S2-185276
5G MoNArch	3GPP-SA2	Solution for updated key issue 1: exposure of analytics to OAM	S2-185278
5G MoNArch	3GPP-SA2	Discussion paper and pCR on V2X slicing	S2-185475
5G MoNArch	3GPP-SA2	Solution for key issue 3: interactions with 5GS NFs/AFs for data collection	S2-185280
5G MoNArch	3GPP-SA2	Update to the general framework for 5G network automation (TR 23.791)	S2-186271
5G MoNArch	3GPP-SA2	TR 23.742: Solution for NF reliability	S2-186151
5G MoNArch	3GPP-SA2	New SID on enhanced support of vertical and LAN services	S2-186182

Project Name	Targeted SDO	Title or short description of contribution related to overall architecture	Status
5G MoNArch	3GPP-SA2	Key Issue: UE-driven analytics sharing mechanisms to 5GC	S2-187264 (S2-186919)
5G MoNArch	3GPP-SA2	UC and KI for KI4 Interactions with OAM for Analytics Exposure	S2-186668
5G MoNArch	3GPP-SA2	New Solution to Key Issue #3: Data Collection by subscription to NFs/AFs	S2-186346
5G MoNArch	3GPP-SA2	Solution: UE-driven analytics sharing	S2-188512 (S2-187903)
5G MoNArch	3GPP-SA2	Solution for Data Collection from OAM using Existing SA5 Services	S2-188263
5G MoNArch	3GPP-SA2	Integration of the 5G System in the TSN network	S2-188459
5G MoNArch	3GPP-SA2	Update to SID: Study of enablers for network automation for 5G	S2-189047
5G MoNArch	3GPP-SA2	TS 23.501 CR0987: CR for TS 23.501 Clarifications NWDAF Discovery and Selection	S2-1902521
5G MoNArch	3GPP-SA2	TS 23.288: Remove the FFS for AF registration during Data Collection procedure	S2-1902398
5G MoNArch	3GPP-SA2	TS 23.288: Analytics exposure to AF via NEF	S2-1902395
5G MoNArch	3GPP-SA2	TS 23.502 CR1060: NEF service for NWDAF analytics	S2-1902524
5G MoNArch	3GPP-SA2	TS 23.501 CR0964: NEF service for NWDAF analytics	S2-1902397
5G MoNArch	3GPP-SA2	23.501 CR1258: Clarifications NWDAF Discovery and Selection	S2-1903964
5G MoNArch	3GPP-SA2	23.502 CR1298: Extensions to NRF Services	S2-1903965
5G MoNArch	3GPP-SA2	P-CR TS 23.288: Clarification of FFS on Analytics Exposure to AFs via NEF	S2-1903966
5G MoNArch	3GPP-SA2	P-CR TS 23.288: Clarifying Flexible AF Registration	S2-1904011
5G MoNArch	3GPP-SA2	TS 23.501 CR1299: Extending Exposure Capability to support Analytics Framework	S2-1903968
5G MoNArch	3GPP-SA2	TS 23.502 CR1300: Updating NEF and NRF Services to Support AF Available Data Registration	S2-1903999

Project Name	Targeted SDO	Title or short description of contribution related to overall architecture	Status
5G TRANSFORMER	3GPP-SA2	New Key Issue: Identify scenarios when Network Slices cannot coexist within a single PLMN	S2-183925 / S2-183923
5G XCAST	3GPP-SA4	Study on V2X Media Handling and Interaction	S4-170715
5G XCAST	3GPP-SA4	SAND for MBMS	S4-170719
5G XCAST	3GPP-SA4	New WID on Usage of CAPIF for xMB API	S4-180283
5G XCAST	3GPP-SA4	New WID on FEC and ROHC activation for GCSE over MBMS (FRASE)	S4-180285
5G XCAST	3GPP-SA4	FS_MBMS_IoT_Timeplan	S4-170579
5G XCAST	3GPP-SA4	Skeleton for TR 26.850 MBMS for IoT v. 0.0.1	S4-170582
5G XCAST	3GPP-SA4	Pseudo-CR-MBMS IoT	S4-170634
5G XCAST	3GPP-SA4	Pseudo-CR on use case for FS_MBMS_IoT	S4-AHI746
5G XCAST	3GPP-SA4	Pseudo-CR on device analysis for FS_MBMS_IoT	S4-AHI747
5G XCAST	3GPP-SA4	Pseudo-CR on overview of LwM2M for FS_MBMS_IoT	S4-AHI748
5G XCAST	3GPP-SA4	Pseudo-CR on MBMS profiles for FS_MBMS_IoT	S4-171006
5G XCAST	3GPP-SA4	Pseudo-Update CoAP overview with block-wise transfer	S4-171205
5G XCAST	3GPP-SA4	Pseudo-Solutions for File Repair procedure using CoAP	S4-171206
5G XCAST	3GPP-SA4	Pseudo-Binary FDT for FS_MBMS_IoT	S4-171207
5G XCAST	3GPP-SA4	pCR 26.881: Performance evaluation of AL-FEC and MCS dimensionning	S4-180128
5G XCAST	3GPP-SA4	pCR 26.850: Binary data formats for MBMS IoT	S4-180180
5G XCAST	3GPP-SA4	pCR 26.850: Solution for announcement during wake-up periods	S4-180075
5G XCAST	3GPP-SA4	pCR 26.850: Solution for announcement of critical data delivery	S4-180181
5G XCAST	3GPP-SA4	pseudo CR to TR26.850 - ASN.1 binary format for reception report message	S4-180534

Project Name	Targeted SDO	Title or short description of contribution related to overall architecture	Status
5G XCAST	3GPP-SA4	ASN.1 binary FDT instance format	S4-180455
5G XCAST	3GPP-SA4	PseudoCR - Solution for service announcement procedures	S4-180537
5G XCAST	3GPP-SA4	pCR Evaluation of file repair solutions	S4-AHI799
5G XCAST	3GPP-SA4	Discussion on the low-end profile for MBMS IoT	S4-AHI800
5G XCAST	3GPP-SA4	pCRs to TR 26.850	S4-AHI801
5G XCAST	3GPP-SA4	Information document on "Trials, Tests and Projects Relating to 5G Broadcast"	SP-180296
5G XCAST	DVB TM-IPI	Adaptive Media Streaming Over IP Multicast	A176
5G XCAST	IETF	Hypertext Transfer Protocol (HTTP) over multicast QUIC	draft-pardue- quic-http- mcast-02
5G XCAST	3GPP	pCR 26.501 consumption report function to UE Media Functions	S4-190471
5G XCAST	3GPP	pCR 26.501 Corrections on consumption report function to UE Media Functions	S4h190830
5G XCAST	3GPP	MCDATA File Distribution support over xMB	S4-180879
SAT5G	3GPP-SA1	Study on using Satellite Access in 5G (TR 22.822)	In-progress
SAT5G	3GPP-SA2	WI "FS_5GSAT_ARCH": Seamless integration of satellite and/or HAPS (High Altitude Platform Station) systems into 5G system; TR 23.737	In-progress
SAT5G	3GPP-SA5	Study on Access Traffic Steering, Switch and Splitting support in the 5G system architecture	In-progress
SAT5G	ETSI TC-SES SCN	WI DTR/SES-00405; Seamless integration of satellite and/or HAPS (High Altitude Platform Station) systems into 5G system; TR 103 611	In-progress

Project Name	Targeted SDO	Title or short description of contribution related to radio and edge architecture	Status
5G CAR	3GPP-RAN2	Mobility challenges for NR V2X platooning	R2-1900606

Project Name	Targeted SDO	Title or short description of contribution related to radio and edge architecture	Status
5G CAR	3GPP-RAN2	Validity area for NR Sidelink resource allocation in V2X communications	R2-1900604
5G CAR	3GPP-RAN2	Discussion on Connection-based versus Connectionless NR Sidelink	R2-1817681
5G CAR	3GPP-RAN2	NR Sidelink resource allocation for V2X communications	R2-1817680
5G CAR	3GPP-RAN2	Various approaches to SL QoS support in NR V2X	R2-1814465
5G ESSENCE	ETSI NFV	Participation in ETSI NFV Plugtests	June 2018 and June 2019
5G ESSENCE	MulteFire	Deployment models	In progress
5G ESSENCE	ETSI ISG MEC	Contribution to ETSI white paper No. 30 MEC in an Enterprise Setting: A Solution Outline	Published
5G ESSENCE	ETSI ISG MEC	Contribution to ETSI white paper No. 20 Developing Software for Multi-Access Edge Computing	Published
5G ESSENCE	ETSI MCX	Participation to MCX Plugtest events, collaborate in preparation of MCX/MCPTT test cases and test architecture	June 2018, Dec 2018-Jan 2019, and March 2019. Planned September 2019.
5G ESSENCE	3GPP SA6	Developing the study on Mission Critical working group CT1	Revision of 3GPP TS 24.281
5G ESSENCE	3GPP SA1	Definition of requirements to support maritime communication services over 3GPP system related to MC services	Revision of 3GPP TS 36.579-2
5G ESSENCE	MEF	Involvement in the MEF 3.0 “5G Implementation project” and an implementation of a PoC demonstrated at the MEF 18 event.	MEF 18 event October-November 2018
5G MoNArch	3GPP-RAN3	Slice information exchange over NG	In progress, R3-180479
5G MoNArch	3GPP-RAN3	Slice information exchange over NG	In progress, R3-181104
5G MoNArch	3GPP-RAN3	Available slice information exchange for NG mobility	In progress, R3-181904

Project Name	Targeted SDO	Title or short description of contribution related to radio and edge architecture	Status
5G MoNArch	3GPP-RAN2	Support for SRB duplication with CA	Accepted, R2-1803233
5G MoNArch	3GPP-RAN2	RLC handling upon duplication deactivation	Accepted, R2-1803318
5G MoNArch	3GPP-RAN3	Slice support of IAB nodes	Noted with consensus, R3-186014
5G TRANSFORMER	ETSI MEC	Instantiating a Network Slice integrating MEC applications, using 3GPP elements	Accepted
5G TRANSFORMER	ETSI MEC	Creation of WI ETSI MEC by 5G-TRANSFORMER	Agreed
5G TRANSFORMER	ETSI MEC	Requirements for MEC Systems with 3rd Parties	Presented
5G XCAST	3GPP-RAN1	New WID on dedicated 5G MBMS for LTE	RP-171603
5G XCAST	3GPP-RAN	New SID Proposal: Study on MBMS for NR	RP-171807
5G XCAST	3GPP-RAN1	New WID on LTE-based 5G Terrestrial Broadcast	RP-181342
5G XCAST	3GPP-RAN	Information document on "Trials, Tests and Projects Relating to 5G Broadcast"	RP-180652
5G XCAST	3GPP-RAN1	Public service broadcaster requirements and background information relevant to LTE-based 5G Terrestrial Broadcast	R1-1810319
5G XCAST	3GPP-RAN1	Scenarios and simulation assumptions for the LTE based 5G terrestrial broadcast gap analysis	R1-1811588
5G XCAST	3GPP-RAN1	Evaluation Results for LTE-Based 5G Terrestrial Broadcast	R1-1812430
5G XCAST	3GPP-RAN1	Evolved Universal Terrestrial Radio Access (E-UTRA); Study on LTE-based 5G terrestrial broadcast	TR36.776
5G XCAST	3GPP-RAN1	Evaluation Results for LTE-Based 5G Terrestrial Broadcast	R1-1902130
5G XCAST	3GPP-RAN1	Network Simulations Regarding the Performance of the CAS	R1-1905330

Project Name	Targeted SDO	Title or short description of contribution related to radio and edge architecture	Status
5G XCAST	3GPP-RAN1	Information For Time Variation Models	R1-1905331
5G XCAST	DVB TM-WIB	Fundamentals of 5G Wideband Broadcasting	TM-WIB0049
5G XCAST	DVB TM-WIB	Implementation Aspects of 5G Wideband Broadcasting	TM-WIB0050
5G XCAST	DVB TM-WIB	Methodology Approach to SIC in Network Simulations	TM-WIB0074
5G XCAST	DVB TM-WIB	Network Simulations with SIC in MFN and SFN	TM-WIB0076
IoRL	ITU-R SM.2422-0, (06/2018)	Visible light for broadband communications	Being presented in late May 2019
IoRL	COST Action CA15104 (IRACON)	Contribution to "Whitepaper on New Localization Methods for 5G"	Presented and published publicly
SAT5G	3GPP-RAN1	Creation of SI "FS_NR_nonterr_nw on NR"; Title: Study on NR to support non-terrestrial networks, TR38.811	RP-171450
SAT5G	ETSI TC-SES SCN	Edge delivery in 5G through satellite multicast	In-progress

Project Name	Targeted SDO	Title or short description of contribution related to core and transport architecture	Status
5G MoNArch	3GPP-SA2	Clarification on key issue: Network slicing for eV2X	S2-183735
5G MoNArch	3GPP-SA2	Updated SID: Study on Enhancement of network slicing	S2-186185
5G MoNArch	3GPP-SA2	LS from FS-eNA to SA5/RAN3	S2-186667
5G MoNArch	3GPP-SA2	Discussion paper on V2X slicing KI	S2-188307
5G MoNArch	3GPP-SA2	Updates to Impacts and Evaluation of Solution 12	S2-1810696
5G MoNArch	3GPP-SA2	Updates to Solution 1 for Network Data Analytics Feedback	S2-1860695

Project Name	Targeted SDO	Title or short description of contribution related to core and transport architecture	Status
5G MoNArch	3GPP-SA2	Solution for KI#2 on Analytics Exposure to AF	S2-1810694
5G MoNArch	3GPP-SA2	Solution to NWDAF assisting traffic routing using MEC information	S2-1810334
5G MoNArch	3GPP-SA2	Updates to Solution 19	S2-1812173
5G MoNArch	3GPP-SA2	Overall Conclusion for Key Issue 4	S2-1812175
5G MoNArch	3GPP-SA2	Updates to Solution 12	S2-1812174
5G MoNArch	3GPP-SA2	Updates to Solution 24	S2-1812172
5G MoNArch	3GPP-SA2	Adding reference to new TS 23.288 in TS 23.502	S2-1901040
5G MoNArch	3GPP-SA2	TS 23.288 skeleton for 5G analytics framework	S2-1901041
5G MoNArch	3GPP-SA2	CR for TS 23.501 based on conclusion of eNA TR 23.791	S2-1901042
5G MoNArch	3GPP-SA2	Adding Selected Solutions #12 from eNA to TS 23.288	S2-1900949
5G MoNArch	3GPP-SA2	Adding Selected Solutions #24 from eNA to TS 23.288	S2-1901024
5G MoNArch	3GPP-SA2	TS 23.288: Update to Data Collection from OAM	S2-1902400
5G MoNArch	3GPP-SA2	P-CR TS 23.288: Setup of Network Map for Data Collection	S2-1903814
5G MoNArch	3GPP-SA2	Update to NF Load Analytics Procedures	S2-1903917
5G MoNArch	3GPP-SA2	Update to Network Performance Analytics Procedures	S2-1903939
5G TANGO	IETF	YANG models for ACTN TE Performance Monitoring Telemetry and Network Autonomics	Draft
5G TRANSFORMER	IETF (CCAMP WG)	draft-ietf-ccamp-microwave-framework-07.txt : 'A framework for Management and Control of microwave and millimeter wave interface parameters'	Draft

Project Name	Targeted SDO	Title or short description of contribution related to core and transport architecture	Status
5G TRANSFORMER	IETF (CCAMP WG)	Finite state machine YANG model augmentation for Transponder Reconfiguration	Draft
5G TRANSFORMER	3GPP SA2	Network slices cannot co-exist within a single PLMN	In-progress
5G TRANSFORMER	3GPP SA2	Clarification on the establishment procedure of proxied MPTCP connections	In-progress
5G TRANSFORMER	3GPP SA2	23.502 CR1053 (Rel-16, 'B'): Slice Specific Secondary Authentication	In-progress
5G TRANSFORMER	3GPP SA2	23.501 CR0951 (Rel-16, 'B'): Slice Specific Secondary Authentication	In-progress
5G TRANSFORMER	IETF (CCAMP WG)	A YANG Data Model for Microwave Topology	Draft
5G TRANSFORMER	IETF (CCAMP WG)	A framework for management and control of microwave and millimeter wave interface parameters	IETF RFC 8432
5G TRANSFORMER	IETF (CCAMP WG)	A YANG Data Model for Microwave Radio Link	Draft
5G XCAST	3GPP-CT4	Support for PWS-IWF	C4-184445
5G XCAST	3GPP-CT1	Removal of Extended Repetition-Period IE for NG-RAN	C1-183005
5G XCAST	3GPP-CT1	PWS in NR -clause 9.1.3.5	C1-181689
5G XCAST	3GPP-CT1	PWS in NR -clause 9.2.0	C1-181690
5G XCAST	3GPP-CT1	PWS in NR -clause 9.2.X	C1-181711
5G XCAST	3GPP-CT1	PWS in NR -clause 9.3.X	C1-181746
5G XCAST	3GPP-CT1	Service Based Interface for 5G system	C1-181647
5G-PICTURE	ITU-T SG15	The Requirements of Mobile-optimized OTN(M-OTN)	in progress
5G-PICTURE	ITU-T SG15	FEC for FlexO-LR interfaces rates of 200 Gbit/s and 400 Gbit/s	in progress
5G-PICTURE	ITU-T SG15	OTU4 long-reach interface	ITU-T Rec. G.709.2

Project Name	Targeted SDO	Title or short description of contribution related to core and transport architecture	Status
blueSpace	ITU-T SG15	Include latency as parameter in Recommendations	Presented to SG
blueSpace	ITU-T SG15	Correlation OTDR for accurate fibre delay measurement	Presented to SG
blueSpace	ITU-T SG15	Modifications to G.697 to include fibre propagation delay	Presented to SG
blueSpace	ITU-T SG15	Proposal to initiate 25G application in G.698.4	Presented to SG
blueSpace	ITU-T SG15	Adding group delay as a parameter into G.671	Presented to SG, draft revised version of G.671 contains proposed changes
METRO-HAUL	ONF	Open Transport API (TAPI)	Complete (v2.1), In Progress (v.2.2)
METRO-HAUL	ITU-T SG15	Variance for Gaussian noise emulation	Approved
METRO-HAUL	ITU-T SG15	EVM and OSNR penalty measurement for draft revised G.698.2	Approved
METRO-HAUL	ITU-T SG15	EVM calculation for G.698.2	Approved
METRO-HAUL	ITU-T SG15	Text proposal for draft G.698.2 defining modulation format of 100G coherent signals	Approved
METRO-HAUL	ITU-T SG15	Equalizer parameters for reference receiver in G.698.2	Approved
METRO-HAUL	IETF	YANG data model for Flexi-Grid Optical Networks	Draft
METRO-HAUL	IETF	YANG models for VN & TE Performance Monitoring Telemetry and Scaling Intent Autonomics	Draft
METRO-HAUL	IETF	YANG data model for Flexi-Grid media-channels	Draft
METRO-HAUL	IETF	A Framework for Enhanced Virtual Private Networks (VPN+) Service	Draft
METRO-HAUL	IETF	Applicability of Abstraction and Control of Traffic Engineered Networks (ACTN) to Network Slicing	Draft

Project Name	Targeted SDO	Title or short description of contribution related to core and transport architecture	Status
METRO-HAUL	OpenROADM	OpenROADM v2 Device Whitepaper	Released

Project Name	Targeted SDO	Title or short description of contributions related to management and orchestration architecture	Status
5G City	ETSI OSM	Initial implementation of vimconn_fos for Eclipse fog05 VIM	Accepted
5G ESSENCE	linuxfoundation	unikraft	In-progress
5G MoNArch	NGMN NWMO	Cross-slice user stories	Completed
5G MoNArch	3GPP SA5	Add Data Analytics Management Service for Network Slice and Network Slice Subnet	Accepted, S5-183560
5G MoNArch	3GPP SA5	Add example of functional management architecture	Accepted, S5-183409
5G MoNArch	ETSI ZSM	Proposal on the overview and architecture of ZSM framework	Accepted, ZSM(18)000236r2
5G MoNArch	ETSI ZSM	Proposed ZSM Architecture Diagram Changes	Accepted, ZSM(18)000325r2
5G MoNArch	ETSI ZSM	Automated discovery of services offered by a management domain	Accepted, ZSM(18)000364r2
5G MoNArch	ETSI ZSM	Definition of integration fabric	Accepted, ZSM(18)000378r1
5G MoNArch	ETSI ZSM	Management service related to network service orchestration	Accepted, ZSM(18)000445
5G MoNArch	ETSI ZSM	Management service related to service performance assurance	Accepted, ZSM(18)000446
5G MoNArch	ETSI ZSM	Add domain performance report service	Accepted, ZSM(18)000450
5G MoNArch	ETSI ZSM	Architecture Diagram Changes	Accepted, ZSM(18)000501
5G MoNArch	ETSI ZSM	Clarify capability of domain orchestration and some clarifications	Accepted, ZSM(18)000442

Project Name	Targeted SDO	Title or short description of contributions related to management and orchestration architecture	Status
5G MoNArch	3GPP SA5	YANG definitions for network slicing NRM	Accepted, S5-185532
5G MoNArch	3GPP SA5	Update the UC and requirements for performance data streaming	Accepted, S5-186429
5G MoNArch	ETSI ENI	Use case on "Elastic resource management and orchestration"	Accepted, ENI(18)000162r1
5G MoNArch	ETSI ENI	Proof of concept on "Elastic network slice management"	Accepted, ENI(18)000175r4
5G MoNArch	3GPP SA5	Update NRM IRP Solution Set to support slice priority	Accepted, S5-187439
5G MoNArch	3GPP SA5	Update NRM root IOCs to support slice priority	Accepted, S5-187370
5G MoNArch	3GPP SA5	Solution for performance data streaming	Accepted, S5-187372
5G MoNArch	ETSI ZSM	Add capabilities to Analytics Service	Accepted, ZSM(18)000596r2
5G MoNArch	ETSI ZSM	Add E2E SLA Management	Accepted, ZSM(18)000601r2
5G MoNArch	ETSI ZSM	Policy management service for E2E	Accepted, ZSM(19)00021
5G MoNArch	ETSI ZSM	Update of the analytics service	Accepted, ZSM(19)000121
5G MoNArch	ETSI ZSM	Discussion on mapping the ZSM002 list of services	Accepted, ZSM(19)000122
5G MoNArch	ETSI ZSM	Update mapping ZSM002 to SA5	Accepted, ZSM(19)000192
5G MoNArch	3GPP SA5	pCR 28.861 Add Multi-dimensional Resource Optimisation	Accepted, S5-193221
5G MoNArch	3GPP SA5	Update NRM requirement to support SBA management	Accepted, S5-193396
5G MoNArch	ETSI ZSM	ZSM002 update of service feasibility check	Accepted, ZSM(19)000195r2
5G MoNArch	ETSI ZSM	ZSM002 Management communication service to solve pub-sub debate	Accepted, ZSM(19)000032r3

Project Name	Targeted SDO	Title or short description of contributions related to management and orchestration architecture	Status
5G MoNArch	ETSI ZSM	Informative examples on ZSM deployment architectures	Accepted, ZSM(19)000203r2
5G TANGO	ETSI OSM	Novel packaging formats aligned with ETSI NFV SOL004	Part of OSM FIVE
5G TANGO	ETSI OSM	Slice Manager allowing the deployment of 5G Network Slices with OSM	Part of OSM FIVE
5G TANGO	IETF	Methodology for VNF Benchmarking Automation	Draft
5G TANGO	ETSI ZSM	Zero-touch Network and Service Management (ZSM); Reference Architecture	Work in progress
5G TANGO	ETSI ZSM	Zero-touch Network and Service Management (ZSM); End to end management and orchestration of network slicing	Work in progress
5G TRANSFORMER	IETF (COMS BoF)	Problem Statement of Common Operation and Management of Network Slicing	Draft
5G TRANSFORMER	ETSI NFV	contribution to ETSI NFV IFA013: NFVIFA(18)000955	In-progress
5G TRANSFORMER	ETSI NFV	Change re quests IFA 013	Approved
5G TRANSFORMER	IRTF (NFVRG)	IPv6-based discovery and association of Virtualization Infrastructure Manager (VIM) and Network Function Virtualization Orchestrator (NFVO)	Draft
5G TRANSFORMER	IETF (COMS BoF)	COMS Architecture	Draft
MATILDA	ETSI OSM	Contributions are related to the support for multi-site deployments as well as the support of runtime policies enforcement.	Work in progress
METRO-HAUL	ETSI OSM	WAN Infrastructure Manager (WIM) Plugin Model	Complete - Release 5
NGPaaS	ETSI NFV	DGR/NFV-IFA029, PaaS with capability supporting container service	Accepted
NGPaaS	ETSI NFV	DGR/NFV-IFA029, Adding container infrastructure management to appendix	Accepted

Project Name	Targeted SDO	Title or short description of contributions related to management and orchestration architecture	Status
SAT5G	3GPP SA5	SI FS_5G_SAT_MO	In-progress
SAT5G	ETSI-SES SCN	WI DTR/SES-00446	In-progress
SLICENET	ETSI ZSM	PoC on zero-touch network slices management and orchestration on a multi-domain environment	In progress
SLICENET	ETSI ENI	PoC proposal for Predictive Fault management of E2E Multi-domain Network Slices	Accepted, ENI