Grant Agreement No.: 814956
Research and Innovation action
Call Topic: ICT-22-2018 EU-China 5G Collaboration

5G HarmoniseD Research and TrIals for serVice Evolution between EU and China

D5.1: First Year Report of 5G Technology and Service Innovations

Version: v1.0

Deliverable type	R (Document, report)
Dissemination level	PU (Public)
Due date	31/08/2019
Submission date	23/08/2019
Lead editor	Olga E. Segou (Orion Innovations P.C.)
Authors	Adrian Quesada Rodriguez, Cédric Crettaz (MI); Sławomir Kukliński, Lechosław Tomaszewski (Orange); Antti Anttonen, Tao Chen (VTT); Nathan Gomes, Shabnam Noor, Phillippos Assimakopoulos, Huiling Zhu, Ignas Laurinavicius (UKent); Ridha Soua, Abdelwahab Boualouach (Uni.lu), Na Yi (UoS), Alexandros Kostopoulos, Maria Belesioti, Anastasia Spiliopoulou (OTE)
Reviewers	Ioannis Chochliouros (OTE), Sławomir Kukliński (Orange)
Work package, Task	WP5 (T5.1, T5.2, T5.3, T5.4)
Keywords	5G, Service Innovations, Slicing, Radio Access, Transport Network, Security, Privacy, Internet of Vehicles

Abstract

5G-DRIVE dedicates Work Package 5 (WP5) to advancements in “5G Technology and Service Innovations”. Its main purpose is to ensure that the 5G test-bed implementations continue to rigorously evolve along the lines of real-world use cases as well as the 5G PPP vision. The work focuses on four distinct aspects, each one forming the core of a dedicated task: (a) Radio Access and Transport Network; (b) Network Virtualisation and Slicing; (c) 5G New Services, and; (d) Security and Privacy.

© 2018 - 2021 5G-DRIVE Consortium Parties

Document revision history

Version	Date	Description of change	List of contributor(s)
v0.0.1	12/02/2019	Document & ToC creation	Olga E. Segou
V0.0.1	14/02/2019	UoS initial contribution to Section 2	Na Yi
V0.0.1	02/2019	Orange input (Section 4)	Sławomir Kuklinski, Lechosław Tomaszewski
V0.0.1	02/2019	MI input (Section 5)	Adrian Quesada Rodriguez; Cédric Crettaz
V0.0.1	02/2019	ORION Input to Section 5	Olga E. Segou
v0.0.2	22/03/2019	Creation of v0.0.2, UKent, VTT and UoS Contribution to Sections 2-3	Tao Chen, Na Yi, Huiling Zhu, Ignas Laurinavicius, Nathan Gomes, Philippos Assimakopoulos
v0.0.2	04/04/2019	ORION Input to Section 5 and appendices	Olga E. Segou
v0.5	05/04/2019	All contributions merged	Olga E. Segou
v0.5.1	11/04/2019	Cross-references and standard elements fixed	Olga E. Segou
v0.5.5	12/04/2019	Delivered for review & GA approval	Olga E. Segou
v.0.5.8	16/04/2019	First round of review completed	Ioannis Chochliouros (reviewer), Olga E. Segou (finalisation)
v.0.6	25/04/2019	Final checking, small corrections	Sławomir Kukliński
v0.7	16/04/2019	Delivered to Coordinator for submission as MS9 report	Olga E. Segou, Uwe Herzog
v0.8	01/07/2019	Sent to partners	Olga E. Segou
v0.9	11/07/2019	OTE inputs to Section 4. Also full conceptual and editorial review of the document	Ioannis Chochliouros, Alexandros Kostopoulos, Maria Belesioti, Anastasia Spiliopoulou
v0.9	14/07/2019	Additional contributions sent by UKent	Philippos Assimakopoulos
v0.9	14/07/2019	Additional conceptual and editorial review of v0.11 of the document	Ioannis Chochliouros
v0.95	15/07/2019	Additional contributions sent by Uni.lu	Ridha Soua
v0.95	16/07/2019	Additional conceptual and editorial review of v0.12 of the document	Ioannis Chochliouros
v0.95	16/07/2019	Added sections: 4.1 (pages 59-70), 4.4 (pages 87-96), 5.5.3 (pages 112-114), 6.3 (pages 131-132), 5.6.1 (pages 115-118) Updated sections: 2.2-2.3 (pages 18-36), 3.3 (pages 37-39)	all
v0.95	26/07/2019	WP leader review and Orange input	Sławomir Kukliński
v0.99	14/08/2019	Editorial review, small corrections and Orange input, update of References	Lechosław Tomaszewski
V0.99	14/08/2019	Sent for final review & GA Approval	Olga E. Segou
V1.0	23/08/2019	Submitted by the Coordinator	Uwe Herzog

Disclaimer

This report contains material which is the copyright of certain 5G-DRIVE Consortium Parties and may not be reproduced or copied without permission.

All 5G-DRIVE Consortium Parties have agreed to publication of this report, the content of which is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License[1].

Neither the 5G-DRIVE Consortium Parties nor the European Commission warrant that the information contained in the Deliverable is capable of use, or that use of the information is free from risk and accept no liability for loss or damage suffered by any person using the information.

CC BY-NC-ND 3.0 License - 2017 5G-DRIVE Consortium Parties.

Acknowledgment (if needed)

This report is based on the milestone MS9 report which was a preliminary version of D5.1, submitted at Month 8 of 5G-DRIVE, to verify the achievement of MS9. The final version of D5.1 provides additional technical detail on all WP5 developments.

[1] *http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en_US*

Executive summary

5G-DRIVE is an innovative 30-month project focused on harmonising research and trials between EU and China on the area of service evolution for enhanced Mobile Broadband (eMBB) and Vehicle-to-Everything (V2X). This document focuses on the work performed with respect to the definition of a novel 5G ecosystem, particularly on:

- radio access and transport networks,
- end-to-end slicing,
- service development and deployment, as well as,
- security and privacy.

5G-DRIVE explores how to provide high performance on radio access by including massive Multiple Input Multiple Output (mMIMO) and virtual resource management under new distributed architecture which includes Distributed Unit (DU) and Central Unit (CU). Powerful computing resources are required to improve the overall radio access performance by adopting the Artificial Intelligence (AI) into MIMO systems and resource allocation. Furthermore, transport network optimisations and the use of Software Defined Networking/Network Function Virtualisation (SDN/NFV) optimisations is considered.

Network slicing enables the creation of parallel virtual telecommunication over a common distributed cloud infrastructure. The main advantage of this approach is the ability of on-demand creation of isolated networking solutions, which are combined or tailored for specific applications and give slice management capabilities to slice tenants. The technical basis for network slicing and slice monitoring is herein discussed.

SDN/NFV service development is extensively discussed. 5G-DRIVE dedicates a lot of effort in ensuring standards compliance. The use of hardware optimisations is extensively discussed as a means to provide high-performance and reliable VNF-based services. An important aspect of this work is the definition of a measurement framework for performance and availability measurements, as this forms the basis for most fault, configuration, administration, performance, security (FCAPS) operations.

© 2018 - 2021 5G-DRIVE Consortium Parties

Table of contents

Executive summary **4**
Table of contents **5**
List of figures **8**
List of tables **11**
Abbreviations **12**
1 Introduction **23**
1.1 Project overview 23
1.2 Scope of the deliverable 23
1.3 Organisation of the document 24
2 Radio access **25**
2.1 Beam squint exploitation for resource allocation in multi-carrier massive MIMO 25
2.2 Modelling of non-stationary channels for MIMO systems 28
2.2.1 Applying the Dent Model with Conventional Water-Filling (CWF) Algorithm 29
2.2.2 Applying channel transformation into visual regions (VRs) and non-visual regions (nVRs) for non-stationary (NS) channels 33
2.3 Deep learning for multi-user detection 33
2.4 Hybrid user association under DU/CU architecture 36
2.4.1 RDP analysis at a BS level 38
2.4.2 Proposed HUA approach 38
3 Transport networks **42**
3.1 Split-RAN segments and requirements 43
3.2 Evolved/Ethernet fronthaul 45
3.3 Analogue fronthaul 47
3.3.1 Up-conversion to mmW frequencies 51
3.3.2 Extensions to arbitrary receiver sampling rates and analogue bandwidths 52
3.4 Function virtualisation and SDN 56
4 Network virtualization and slicing interoperability **59**
4.1 Network slicing concept 59
4.1.1 General introductory framework 59
4.1.2 Network Slicing in the 5G digital era 59
4.1.3 Actual trends for further growth 63
4.1.4 Discussion upon the network slicing-related conceptual approach 66
4.1.5 Network slicing status 70
4.1.6 Standardization of network slicing 70
4.1.7 Network slicing-related research activities 72
4.2 Open issues 74

4.2.1 Addressing highly differentiated slicing needs and contexts ... 74
4.2.2 Architecture-related issues ... 74
4.2.3 Slice description and selection issues ... 75
4.2.4 Service to slice relationship ... 75
4.2.5 Data plane issues ... 75
4.2.6 RAN issues ... 76
4.2.7 Slice isolation issues ... 76
4.2.8 Management issues ... 77
4.2.9 Orchestration issues ... 77
4.2.10 Multi-domain slicing-related issues ... 77
4.2.11 Business models of network slicing ... 78
4.2.12 Network slicing and legal issues ... 78
4.3 New KPIs for network slicing ... 78
4.3.1 Background ... 79
4.3.2 Network slicing KPIs set proposal ... 81
4.3.3 Network slicing KPIs computation in the NFV MANO case ... 85
4.4 Dynamic Network Slicing ... 87
4.4.1 Dynamic Network slicing as Enabler for End-to-End Performance and for a fully Programmable Network ... 89
4.4.2 NFV and SDN as Enablers for Network Slicing ... 90
4.4.3 Management and Orchestrating of Dynamic End-to-End Network Slices ... 90
4.4.4 Dynamic End-to-End Network Slicing for optimizing the entire 5G Network ... 92
4.4.5 Potential Future Trends in the scope of Dynamic Network Slicing ... 93
4.4.6 Potential “Go-to-Market” Strategy and Regulatory Aspects ... 95
4.5 Multi-provider infrastructure extensions for MANO ... 96
5 Novel services and cybersecurity ... 98
5.1 ETSI NFV reference architecture ... 98
5.2 The VNF domain ... 99
5.2.1 VNF & NS composition ... 99
5.2.2 VNF Interfaces ... 99
5.3 The NFVI domain ... 100
5.3.1 ORION local NFVI ... 100
5.3.2 High-performance VNF development with Enhanced Platform Awareness ... 101
5.4 The MANO domain ... 106
5.4.1 General architecture ... 106
5.4.2 Suggested MANO stack ... 106
5.5 Instantiation and lifecycle management ... 107
5.5.1 Images, descriptors and records ... 107

5.5.2 Instantiation and configuration process 108
5.5.3 VNF Optimisations with EPA 112
5.6 Integration, testing and assurance 114
5.6.1 Integration plan 115
5.6.2 Tools and KPIs 119
5.7 Security and privacy aspects 123
5.7.1 Security in 5G V2X 123
5.7.2 Privacy in 5G V2X 126
6 Conclusions 130
6.1 Final remarks 130
6.2 Future work and important considerations 131
6.3 Contributions to WP3 & WP4 131
6.3.1 Network slicing 131
6.3.2 Service evolution and testing 131
6.3.3 Channel modelling for MIMO systems 132
6.3.4 Evolved Ethernet network: latency and SDN processing 132
7 References 133
Appendix A OpenStack and OSM installation and configuration 138

List of figures

Figure 1: Beam squint in an array with 128 antenna elements, the antenna spacing is $d = \lambda c/2$, where λc is the wavelength for the centre frequency. 26

Figure 2: Channel frequency response for a system with 128 antenna elements, antenna spacing $\lambda c/2$ and the expected beam direction at 38°. 27

Figure 3: Maximum achievable capacity for an increasing number of antennas 28

Figure 4: Capacity Improvement Multi-User *vs.* Single User. SNR for both curves is chosen to be 12dB, N_{RF}=3..... 28

Figure 5: Dent channel model incorporating the Li-Huang model as M independent and orthogonal scatterers Akp and Akq, which are modeled by applying the Walsh-Hadamard sequences. 30

Figure 6: Received signal strength indicator (RSSI) of the Dent channel model at $28GHz$. 30

Figure 7: Single-sided spectrum (Doppler spread) of the Dent channel model at $28GHz$, indicating a spread of up to 150 kHz. 30

Figure 8: Normalized autocorrelation of the first waveform (k=1)..... 31

Figure 9: Normalized cross-correlation between the waveforms. 31

Figure 10: Quantized RSSI output at the receiver set to 6dB. 32

Figure 11: Spectral Efficiency with CWF equalization for a non-stationary channel. 32

Figure 12: Block diagram of equalisation-and-learning for multiuser-SIMO detection. 34

Figure 13: Block diagram of direct-learning for multiuser-SIMO detection. 34

Figure 14: Average bit-error-rate performance of the equalisation-and-learning paradigm. 35

Figure 15: Average bit-error-rate performance of the directly-learning paradigm. 36

Figure 16: Proposed HUA approach. 39

Figure 17: Pseudocode of the HUA approach. 40

Figure 18: Comparison of RDP for different UA methods. 40

Figure 19: Comparison of tier association probability for different UA methods. 41

Figure 20: Comparison of SINR outage probability for different UA methods. 41

Figure 21: New transport interfaces defined by 3GPP for 5G networks. 42

Figure 22: Visualisation of 5G transport network. 43

Figure 23: HLS/F1 interface in 3GPP RAN protocol stack. 44

Figure 24: Possible LLS/F2 interfaces for 5G networks. 44

Figure 25: MAC-PHY (Option 6) Lower-layer Split Implementation with Ethernet fronthaul. 46

Figure 26: MAC-PHY (Option 6) Lower-layer Split throughput measurement. 46

Figure 27: DSP-assisted analogue fronthaul. 47

Figure 28: Conceptual view of the techniques used in the DSP-assisted RoF system. 48

Figure 29: The process of channel creation/placement at the input of the single-IFFT stage. 49

Figure 30: Conceptual view of the de-multiplexing process at the receiver. 49

Figure 31: Experimental set-up. 50

© 2018 - 2021 5G-DRIVE Consortium Parties

Figure 32: 64-channel multiplex: A. Transmit spectrum; B. Received one-sided spectrum; C. Received signal constellation for a channel selection. 51

Figure 33: mmW up/down-conversion set-up. 52

Figure 34: Received channel spectra and constellations following mmW up/down-conversion. 52

Figure 35: Incorporation of a THA with minimal analogue-domain filtering resulting in the ability to employ receivers with arbitrary low sampling rates and analogue bandwidths. 53

Figure 36: The model used for the receiver side front-end, now incorporating a THA. 54

Figure 37: 120-channel multiplex with mixed 5G channel bandwidths/numerologies: Two channels per NZ with one channel having a bandwidth of 20 MHz and a subcarrier spacing of 30 kHz and the other a bandwidth of 80 MHz and subcarrier spacing of 60 kHz. The resulting data rate is 21.6 Gbps (16-QAM)/32.4 Gbps (64-QAM). 54

Figure 38: Received signal spectrum and constellation for a selection of channels, one form each numerology, from the 120-channel multiplex. 55

Figure 39: A 240-ch multiplex with mixed 5G channel bandwidths/numerologies and different modulation schemes: Two channels per NZ with one channel having a bandwidth of 20 MHz and a subcarrier spacing of 30 kHz and the other a bandwidth of 80 MHz and subcarrier spacing of 60 kHz. The resulting data rate is 49.7 Gbps. 55

Figure 40: Received signal spectrum and constellation for a selection of channels, one form each numerology, from the 240-channel multiplex following up/down-conversion to/from 60 GHz. 56

Figure 41: SDN in an Ethernet fronthaul. 57

Figure 42: SDN reconfiguration possibilities for an Ethernet fronthaul. 57

Figure 43: SDN traffic steering within an evolved fronthaul (MAC/PHY LTE functional split). 58

Figure 44: The concept of network slicing [Source: ITU-T Recommendation Y.3112 (12/2008)]..... 67

Figure 45: Network slicing conceptual outline. 68

Figure 46: Multi-domain management and orchestration architecture for end-to-end KPI calculation. 85

Figure 47: 5G network slicing addressing a variety of industry and multi-tenancy use cases with multiple independent instances on one physical network. [Source: Nokia (2016): “Dynamic End-to-End Network Slicing for 5G- White Paper”]. 88

Figure 48: ETSI NFV Standard Architecture and basic technology mapping. 98

Figure 49: Example architecture of a Network Service with multiple VNFs. 99

Figure 50: Types of VNF interfaces. 100

Figure 51: ORION Athens NFVI-PoP (including additional infrastructure). 101

Figure 52: (a) EPA benefits for fine-grained development; (b) Support for advanced features must be configured throughout the infrastructure. 104

Figure 53: ETSI MANO information elements. Descriptors are provided by the VNF developers, while records are created upon instantiation of a network service. 108

Figure 54: Sequence diagram of the ETSI MANO on-boarding process. 109

Figure 55: Sequence diagram of the ETSI MANO enabling process. 109

Figure 56: Sequence diagram for the instantiation of a 5G-DRIVE service. 111

Figure 57: Excerpt from the OSM Information Model for the VNF descriptor, showing definition of EPA features, marking some potential features for the case of the 5G ESSENCE project. 112

Figure 58: Node_exporter use for KPI monitoring from a VM. 119
Figure 59: Software defined vehicular network architecture 124
Figure 60: OpenStack logical architecture, showing prioritisation of services for the 5G-DRIVE..... 139

List of tables

Table 1: Comparison between RAN split point requirements, adapted from [15]. 45

Table 2: Latency measurements for MAC-PHY split Ethernet fronthaul. 47

Table 3: IMT-2020 Focus Group reports. 71

Table 4: Comparison between Quality Assessment Approaches. 79

Table 5: Calculating KPIs using MANO. 86

Table 6: Enhanced Platform Awareness NFVI features available for OpenStack (as of version Rocky – the list is non-exhaustive). 102

Table 7: Performance enhancements per Compute feature, published by Intel. 103

Table 8: Example listing the NFVI & VIM configurations necessary (per feature). 105

Table 9: Example of the configuration files that need to be edited in OpenStack (per feature). 105

Table 10: Standard VNF primitives available in OSM (and applied as Juju actions). 110

Table 11: Integration plan for T5.3 services with telemetry. 116

Table 12: KPIs collected by Prometheus through use of node_exporter for Linux-based services. 119

Table 13: Tools for NFVI & NFV performance & availability tests. 121

Table 14: Availability characteristics. 122

Table 15: EU and China regulatory framework. 127

Table 16: Additional OSM components that can be enabled during installation. 140

© 2018 - 2021 5G-DRIVE Consortium Parties

Abbreviations

3GPP	The Third Generation Partnership Project
4G	The Fourth Generation of Mobile Communications
5G	The Fifth Generation of Mobile Communications
5GC	5G Core network
5GPPP	Fifth Generation (5G) Public Private Partnership
ADC	Analogue to Digital Converter
AES	Advanced Encryption Standard
AES-NI	Advanced Encryption Standard - New Instructions
AI	Artificial Intelligence
AMF	Access and Mobility Function
AoA	Angle of Arrival
AoD	Angle of Departure
API	Application Programming Interface
app	Application
AR	Augmented Reality
ARIB	Association of Radio Industries and Businesses
ARQ	Automatic Repeat Request
ATIS	Alliance for Telecommunications Industry Solutions Deployments
AVX	Advanced Vector Extensions
AWG	Arbitrary Waveform Generator
BBU	Baseband Unit
BER	Bit-Error Rate
BIOS	Basic Input Output System
BOF	Birds of a Feather
BPSK	Binary Phase Shift Keying
BS	Base Station
BSS	Business Support System
BTS	Base Transceiver Station
CAPEX	Capital Expenditures
CCSA	China Communications Standards Association
CCU	Central Control Unit
CH, ch	Channel
CH	Cluster Head
CLI	Command-Line Interface

© 2018 - 2021 5G-DRIVE Consortium Parties

CM	Configuration Management
CN	Core Network
ComRo	Computing Resources Underutilisation
ComRu	Computing Resources Underutilisation
ConRo	Connectivity Resources Overutilisation
ConRu	Connectivity Resources Underutilisation
CoS	Class of Service
CP	Connection Point
CP	Control Plane
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
CSI-RSRP	Channel State Information Reference Signal Received Power
CSI	Channel State Information
CSIR	Channel State Information at the Receiver Side
CSP	Communications Service Provider
CU	Central Unit
CUPS	Control Plane and User Plane Splitting
CWF	Conventional Water-Filling
CWL	Continuous Wave Laser
D2D	Device-to-Device
DAC	Digital to Analog Converter
DAS	Distributed Antenna System
DCI	Downlink Control Information
DDR	Double Data Rate
DDR3	Double Data Rate 3
DÉCOR	Dedicated Core
DHCP	Dynamic Host Configuration Protocol
DiffServ	Differentiated Service
DL	Direct Learning
DLSCH	Downlink Shared Channel
DMT	Discrete Multi-Tone
DNN	Deep Neural Network
DoA	Description of Action
DP	Data Plane
DPDK	Data Plane Development Kit
DSP	Digital Signal processor

© 2018 - 2021 5G-DRIVE Consortium Parties

DU	Distributed Unit
DWDM	Dense Wavelength Division Multiplexing
DUA	Decentralised User Association
E2E	End-to-End
E-UTRA	Evolved UMTS Terrestrial Radio Access
EDFA	Erbium-Doped Fiber Amplifier
EEA	European Economic Area
EL	Equalisation-and-Learning
EM	Elemental Manager
eMBB	Enhanced Mobile Broadband
EMS	Element Manager System
eNB	evolved NodeB
EPA	Enhanced Platform Awareness
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
EU	European Union
EVM	Error Vector Magnitude
FaaS	Function as-a-Service
FAB	Fulfilment, Assurance and Billing
FCAPS	Fault, Configuration, Administration, Performance and Security
FEC	Forward Error Correction
FFT	Fast-Fourier Transform
FM	Fault Management
FPGA	Field Programmable Gate Array
FSB	Front Side Bus
FW	Firewall, Fire Wall
GA	Grant Agreement
GDPR	General Data Protection Regulation
GHz	Giga Hertz
GPRS	General Packet Radio Service
GTP	GPRS Tunnelling Protocol
GPS	Global Positioning System
GPU	Graphics Processing Unit
GR	Group Report
GS	Group Specification
GSM	Global System for Mobile Communications

© 2018 - 2021 5G-DRIVE Consortium Parties

GSMA	GSM Association
GUI	Graphical User Interface
H2020	Horizon 2020
HARQ	Hybrid Automatic Repeat Request
HD	High Definition
HetNet	Heterogeneous Network
HLS	Higher Layer Split
HP	Hyper-Treading
HPE	Hewlett Packard Enterprise
HTTP, http	HyperText Transfer Protocol
HUA	Hybrid User Association
HW	Hardware
I2I	Infrastructure-to-Infrastructure
IaaS	Infrastructure as-a-Service
IQ	In-Phase and Quadrature
ICT	Information and Communication Technology
ID, id	Identifier
IDS	Intrusion Detection System
IEEE	Institute of Electrical and Electronic Engineers
IETF	Internet Engineering Task Force
IF	Intermediate Frequency
IFA	International Financial Architecture
IFFT	Inverse Fast-Fourier Transform
ILS	Integer Least-Squares
IMT	International Mobile Telecommunications
IntServ	Integrated Service
IO, I/O	Input / Output
IOMMU	Input Output Memory Management Unit
iOS	iPhone Operating System
IoT	Internet of Things
IoV	Internet of Vehicles
IP	Internet Protocol
IPR	Intellectual Property Rights
IPVS	IP Virtual Server
iSCSI	Internet Small Computer Systems interface
ISI	Inter-Symbol Interference

© 2018 - 2021 5G-DRIVE Consortium Parties

ISG	Industry Specifications Group
ISP	Internet Service Provider
IT	Information Technology
ITS	Intelligent Transportation System
ITU	International Telecommunications Union
ITU-R	International Telecommunications Union - Radiocommunication Sector
ITU-T	International Telecommunications Union - Telecommunication Standardization Sector
JHA	Justice and Home Affairs
KPI	Key Performance Indicator
KQI	Key Quality Indicator
KVM	Kernel-based Virtual Machine
LAN	Local Area Network
LLS	Lower Layer Split
LMMSE	Linear Minimum Mean-Square Error
LO	Local Oscillator
LR	Lattice-Reduction
LTE	Long Term Evolution
MaaS	Management as-a-Service
MAC	Medium Access Control
MANO	Management and Orchestration
MBB	Mobile Broadband
MC	Mission-Critical
MCPTT	Mission-Critical Push-To-Talk
MEC	Mobile Edge Computing
MEC	Multi-access Edge Computing
MemRo	Memory Resources Overutilisation
MemRu	Memory Resources Underutilisation
MF	Matched Filter
mgnt	Management
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MME	Mobility Management Entity
mMIMO	Massive Multiple Input Multiple Output
MMSE	Minimum Mean-Square Error
mMTC	Massive Machine Type Communications

© 2018 - 2021 5G-DRIVE Consortium Parties

MMU	Memory Management Unit
mmW	millimeter wave
MTBF	Mean-Time-Between-Failures
MTC	Machine Type Communications
MTTR	Mean-Time-To-Repair
MU	Multi-User
MUD	Multi-User Detection
MU-MIMO	Multiuser Multiple-Input Multiple-Output
MU-SIMO	Multiuser Single-Input Multiple-Output
MVNO	Mobile Virtual Network Operator
MZM	Mach-Zenhder Modulator
N/A	Not Available
NaaS	Network as-a-Service
NAT	Network Address Translation
NB	Narrow Band
NB-IoT	Narrow Band Internet of Things
NCU	Network Capability Unit
NF	Network Function
NFS	Network File System
NFV	Network Function Virtualisation
NFVI	Network Function Virtualisation Infrastructure
NFVO	Network Function Virtualisation Orchestrator
NG	Next Generation
NGMN	Next Generation Mobile Networks
NGP	Next Generation Protocol
NIaaS	Network Infrastructure as-a-Service
NIC	Network Interface Card
NIC	Network Interface Controller
NMS	Network Management System
NPaaS	Network Platform as-a-Service
NR	New Radio
NRT	Near Real Time
NS	Network Service
NS	Non-Stationary
NSaaS	Network Slicing as a Service
NSD	Network Service Descriptor

© 2018 - 2021 5G-DRIVE Consortium Parties

NTP	Network Time Protocol
NUMA	Non-Uniform Memory Access
nVR	Non-Visual Region
NZ	Nyquist Zone
OAI	Open Air interface
OBSAI	Open Base Station Architecture Initiative
OComRo	Overall Computing Resources Overutilisation
OComRu	Overall Computing Resources Underutilisation
OConRo	Overall Connectivity Resources Overutilisation
OConRu	Overall Connectivity Resources Underutilisation
ODL	OpenDaylight
OFDM	Orthogonal Frequency Division Multiplexing
OMemRo	Overall Memory Resources Overutilisation
OMemRu	Overall Memory Resources Underutilisation
OS	Open Source
OSC	Oscilloscope
OSM	Open Source MANO
OSS	Operations Support System
OTA	Over-The-Air
OVS	Open vSwitch
P2P	Peer-to-Peer
PaaS	Platform as-a-Service
PC	Personal Computer
PCI	Peripheral Component Interconnect
PCIe	Peripheral Component Interconnect Express
PCP	Pre-Commercial Procurement
PCS	Personal Communications Service
PDCP	Packet Data Convergence Protocol
PHY	Physical Layer
PLMN	Public Land Mobile Network
PNF	Physical Network Function
PNFD	Physical Network Function Descriptor
PoC	Proof-of-Concept
PoP	Point-of-Presence
PPP	Poisson Point Process
PRC	People’s Republic of China

© 2018 - 2021 5G-DRIVE Consortium Parties

PSK	Phase Shift Keying
PTP	Precision Time Protocol
QAM	Quadrature Amplitude Modulation
QAT	Quick Assisted Technology
QEMU	Quick Emulator
QoE	Quality of Experience
QoS	Quality of Service
RAM	Random Access Memory
RAN	Radio Access Network
RAR	Random-Access Request
RAT	Radio Access Technology
RAU	Remote Antenna Unit
RB	Radio Bearer
RCE	Relative Constellation Error
RDP	Resource Depletion Probability
RAR	Random-Access Request
RE	Radio Equipment
REC	Radio Equipment Controller
RET	Reconfiguration Execution Time
RF	Radio Frequency
RIA	Research and Innovation Action
RID	Requester ID
RLC	Radio Link layer Control
RNC	Radio Network Controller
RO	Resource Orchestration
RoF	Radio over Fiber
RRC	Radio Resource Control
RRH	Remote Radio Head
RRU	Remote Radio Unit
RSDNC	RSU-SDN Controller
RSRP	Reference Signal Received Power
RSSI	Received Signal Strength Indicator
RSU	Road-Side Unit
RT	Real Time
Rx	Reception
S-NSSAI	Single Network Selection Assistance Information

© 2018 - 2021 5G-DRIVE Consortium Parties

SS-RSRP	Synchronization Signal Reference Signal Received Power
SS-SINR	Synchronization Signal Signal to Interference plus Noise Ratio
SaaS	Software as-a-Service
SAS	Statistical Analysis System
SBA	Service-Based Architecture
SC	Small Cell
SCM	Subcarrier Multiplexing
SCSI	Small Computer Systems interface
SD	Slice Differentiation
SD	Slice Differentiator
SD	Software Defined
SD	Sphere Decoding
SDAP	Service Data Adaptation Protocol
SDMN	Software Defined Mobile Network
SDN	Software Defined Networking
SDO	Standards Developing Organisation
SDR	Software Defined Radio
SDT	Slice Deployment Time
SDTS	Slice Deployment Time Scalability
SECaaS	Security as-a-Service
SER	Symbol Error Rate
SG	Study Group
SGD	Stochastic Gradient Descent
SGX	Software Guard Extensions
SGW	Serving Gateway
SI	System Information
SIC	Soft Interference Cancellation
SIM	Subscriber Identity Module
SIMO	Single-Input, Multiple-Output
SINR	Signal to Interference plus Noise Ratio
SLA	Service Level Agreement
SMF	Single Mode Fiber
SMS	Short Message Service
SNR	Signal-to-Noise Ratio
SR-IOV	Single Root I/O Virtualisation
SSB	Single-SideBand

© 2018 - 2021 5G-DRIVE Consortium Parties

SST	Slice Service Type
STD	Standard Deviation
STT	Slice Termination Time
SW	Software
SWA	Software Assurance
TBD, tbd	To Be Defined
TCP	Transmission Control Protocol
THA	Track-and-Hold Amplifier
ToC	Table of Contents
TPM	Trusted Platform Module
TR	Technical Report
TS	Technical Specification
TSDSI	Telecommunications Standards Development Society India
TSN	Time-Sensitive Networking
TTA	Telecommunications Technology Association
TTC	Telecommunications Technology Committee
TTD	True Time Delay
Tx	Transmission
TXT	Trusted Execution Technology
UA	Unified Architecture
UA	User Association
UC	Use Case
UDP	User Datagram Protocol
UE	User Equipment
UHD	Ultra-High Definition
UI	User Interface
UMTS	Universal Mobile Telecommunications System
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communications
UT	User Terminal
UTRA	UMTS Terrestrial Radio Access
UTRAN	UMTS Terrestrial Radio Access Network
V2I	Vehicle-to-Infrastructure
V2N	Vehicle-to-Network
V2V	Vehicle-to-Vehicle

© 2018 - 2021 5G-DRIVE Consortium Parties

V2X	Vehicle-to-Everything
VCA	Vehicle Certification Agency
vCU	virtual Central Unit
vDU	virtual Distributed Unit
vEPC	virtual Evolved Packet Core
VIM	Virtualised Infrastructure Manager
VL	Virtual Link
VLAN	Virtual Local Area Network
VLD	Virtual Link Descriptor
VLPZ	Vehicular Location Privacy Zone
VM	Virtual Machine
VNF	Virtual Network Function
VNFaaS	VNF as-a-Service
VNFC	Virtual Network Function Component
VNFD	VNF Descriptor
VNFFG	VNF Forwarding Graph
VNFFGD	VNF Forwarding Graph Descriptor
VNFM	Virtual Network Function Manager
VNFO	Virtual Network Function Orchestrator
VPI	Virtual Photonics Incorporated
VR	Virtual Reality
VR	Visual Region
VSDNC	Vehicular-SDN Controller
WAVE	Wireless Access in Vehicular Environments
WiFi, Wi-Fi	Wireless Fidelity
WWW, www	World Wide Web
WG	Working Group
WP	Work Package
XML, xml	eXtensible Markup Language
XFS	Extended File System
ZF	Zero-forcing
ZFS	Zettabyte File System

© 2018 - 2021 5G-DRIVE Consortium Parties

1 Introduction

1.1 Project overview

5G-DRIVE is an innovative 30-month project focused on harmonising research and trials between EU and China on the area of service evolution for enhanced Mobile Broadband (eMBB) and Vehicle-to-Everything (V2X) communications. The 5G-DRIVE project is supported by a consortium comprising of 17 partners from EU academia, industry and commercial areas.

The project's objectives are structured into three main areas, that is: technical, regulatory and business objectives. These are listed, in brief, as follows:

Technical objectives

- Build pre-commercial end-to-end (E2E) testbeds in two cities with sufficient coverage to perform extensive eMBB and Internet of Vehicles (IoV) trials. Joint test specifications will be defined through the collaborative agreement with the Chinese project.
- Develop and trial key 5G technologies and services, including but not limited to: massive multi-input multi-output (MIMO) at 3.5 GHz; end-to-end network slicing; mobile edge computing (MEC) for low latency services and V2X; software-defined networking (SDN) for transport and core network, and; network and terminal security.
- Develop and trial cross-domain network slicing techniques across two regions for new services.
- Demonstrate IoV services using Vehicle-to-Network (V2N) and Vehicle-to-Vehicle (V2V) communications operating at 3.5 GHz and 5.9 GHz, respectively.
- Analyse potential system interoperability issues identified during the trials in both regions and to provide joint reports, white papers and recommendations to "address" them, accordingly.
- Submit joint contributions to 3GPP and other 5G standardisation bodies regarding the "key" 5G technologies developed and evaluated in the project.

Regulatory objectives

- Evaluate spectrum usage at 3.5 GHz for indoor and outdoor environments in selected trial sites and provide joint evaluation reports and recommendations on 5G key spectrum bands in Europe and China.
- Investigate regulatory issues regarding the deployment of V2X technologies (i.e. coexistence in the 5.9 GHz band) and provide joint reports.

Business objectives

- Investigate and promote 5G business potential through joint development of 5G use cases and applications.
- Strengthen industrial 5G cooperation between the EU and China.
- Promote early 5G market adoption through joint demonstrations in large showcasing events.

1.2 Scope of the deliverable

The 5G-DRIVE context dedicates an entire Work Package (WP5) to improvements in "5G Technology and Service Innovations". The main purpose of this specific WP5 is to develop and document key 5G improvements to support real operational scenarios, in terms of expected functionalities, as well as scalability and performance characteristics. Specifically:

© 2018 - 2021 5G-DRIVE Consortium Parties

- **Radio Access and Transport Network** (T5.1) centres around RAN improvements, Distributed Massive MIMO and data transport beyond 5G mobile communications.
- **Network Virtualisation and Slicing** (T5.2) focuses on the development of a context-aware, multi-level control plane and E2E slicing.
- **5G New Services** (T5.3) address novelties required at the service level in order to flexibly provision, replace, migrate and troubleshoot network functions, as well as monitor their SLA assurance.
- **Security and Privacy aspects of 5G and the Internet of Vehicles** (T5.4) tackles security and privacy challenges within the complex 5G ecosystem.

1.3 Organisation of the document

The document is organised as follows:

- **Section 1** (current section) serves as an overall introduction to the document and discusses the scope of WP5.
- **Section 2** discusses Radio Access Technologies (T5.1), including massive MIMO.
- **Section 3** focuses on the Transport Network (T5.1).
- **Section 4** focuses on Network Virtualisation and Slicing (T5.2) including aspects of MANO operations.
- **Section 5** provides an overview of T5.3/T5.4, focusing upon the deployment of high-performance network services and upon security/privacy considerations.
- **Section 7** summarises the key points and conclusions drawn from the work performed during M1-M8 and provides the next steps towards D5.1.
- **Appendix A** provides the process for a test configuration and optimisation of a service ecosystem.

© 2018 - 2021 5G-DRIVE Consortium Parties

2 Radio access

In 5G communications, new radio access techniques are developed to enhance the overall network. In the 5G-DRIVE framework, the key radio access technologies to be explored to provide high performance include massive Multiple Input Multiple Output (mMIMO) and virtual resource management under a new distributed architecture which includes Distributed Unit (DU) and Central Unit (CU). Also, the powerful computing resource is considered to improve the overall radio access performance by adopting the Artificial Intelligence (AI) into MIMO systems and resource allocation. In this section, solutions to three issues in massive MIMO systems and the new distributed architecture are discussed and evaluated by simulations. In order to investigate the precoding schemes in MIMO systems, the channel modelling is also important. According to the test results, the mobile channels in MIMO systems may not be stationary. Therefore, how to model non-stationary channels is also investigated.

2.1 Beam squint exploitation for resource allocation in multi-carrier massive MIMO

In massive multiple input multiple output (MIMO) systems, the antenna array could include a large number of antenna elements, for example, hundreds of antenna elements. With phased arrays configured for the antenna elements, narrow beams can be formed with large antenna gains towards users, and there are different configurations available that meet beamforming needs. A system with an RF chain dedicated to each antenna is known as a digital beamformer[2], which has the advantage of most flexibility, allowing it to reach the highest capacity [1]. The main challenge to digital beamforming is that the cost of the implementation is extremely high. Comparatively, an analogue beamformer is the easiest to implement, as multiple antenna elements could share one RF chain, while this configuration would have the least amount of flexibility and relatively low capacity[3]. A configuration incorporating an analogue precoder of phase-shifters and a digital precoder is known as Hybrid beamforming, which works as a compromise between achievable performance and implementation cost[4].

Typically, the analogue beamformer is configured with the phased array. Each branch of the phased array generally has the same phase shift for all frequencies within the communication band. The phase shift is usually approximated for a fixed time delay for narrowband signals. However, the approximation would break down for signals with wide bandwidth when the Angle of Arrival[5] (AoA) or Angle of Departure (AoD) is not zero. In the context of narrowband assumption, if the carrier

---

[2] Beamforming or spatial filtering is a signal processing technique used in sensor arrays for directional signal transmission or reception. This is achieved by combining elements in an antenna array in such a way that signals at particular angles experience constructive interference while others experience destructive interference. Beamforming can be used at both the transmitting and receiving ends in order to achieve spatial selectivity. The improvement compared with omnidirectional reception/transmission is known as the directivity of the array. For further information also see, *among others*: Wong, A.C.C. (1989): *Considerations in Digital Beamforming Design*. In Proceedings of the IET Colloquium of Multiple Beam Antennas and Beamformers, London, UK, November 21, 1989.

[3] For more details about differences between digital and analogue beamforming also see, *among others*: *http://www.rfwireless-world.com/Terminology/Analog-Beamforming-vs-Digital-Beamforming.html*

[4] Also see: A.F. Molish, V.V. Ratnam, S. Han, Z. Li, at *al.* (2017): Hybrid Beamforming for Massive MIMO - A Survey. *IEEE communications Magazine,* vol. 55, no. 9, pp. 134-141.

[5] Angle of arrival (AoA) measurement is a method for determining the direction of propagation of a radio-frequency wave incident on an antenna array or determined from maximum signal strength during antenna rotation. The AoA determines the direction by measuring the time difference of arrival at individual elements of the array - from these delays the AoA can be calculated.

frequency is f_c, in order to point the beam to a direction with angle of θ, the analogue beamforming weight for each antenna element, w_m, is calculated according to the value of f_c and θ, so that a beamforming vector $\boldsymbol{w}=[w_1,w_2,\cdots,w_M]$ is obtained, where M is the number of antenna elements. For a narrowband system, with the same beamforming vector, the channel gain could be approximated to be the same for all the frequency. However, in a wideband system with bandwidth B, the channel gain at an operating frequency f, where $f\in[f_c-B/2, f_c+B/2]$, will be changed by a factor of $a(\theta,f)$, which varies with the AoA/AoD angle θ and operating frequency f [2].

Figure 1 demonstrates the frequency response at different AoA/AoD, when a system with a fixed analogue beamforming vector operating at three different frequencies, f_1=65 GHz, f_2=81 GHz and the centre frequency f_c =73 GHz. In this simulation, the number of antenna elements is 128. The bandwidth is B = 16 GHz = $0.2f_c$. The expected beam direction is θ=30°. It can be seen from this figure that when operating at a frequency other than the centre frequency, the beam direction is different from the expected direction achieved at the centre frequency, which is referred to as the beam squint[6].

Figure 2 illustrates channel frequency response for a system with a carrier/centre frequency f_c at 73 GHz and AoA/AoD at 38°. In the figure, $\xi= f/f_c$ indicates the ratio of the operating frequency to the carrier/centre frequency. From this figure, it can be seen that, when the channel bandwidth B is relatively large, the channel gain is not flat for the whole bandwidth, and the effect of operating frequency cannot be ignored.

It can be seen from both Figure 1 and Figure 2 that the effect of beam squint on the system performance could be significant in a system with relatively large bandwidth. With the increase in the number of antennas, beam squint could be more severe. It is also affected by the number of antennas. which could be a serious issue to be addressed in millimetre wave massive MIMO systems.

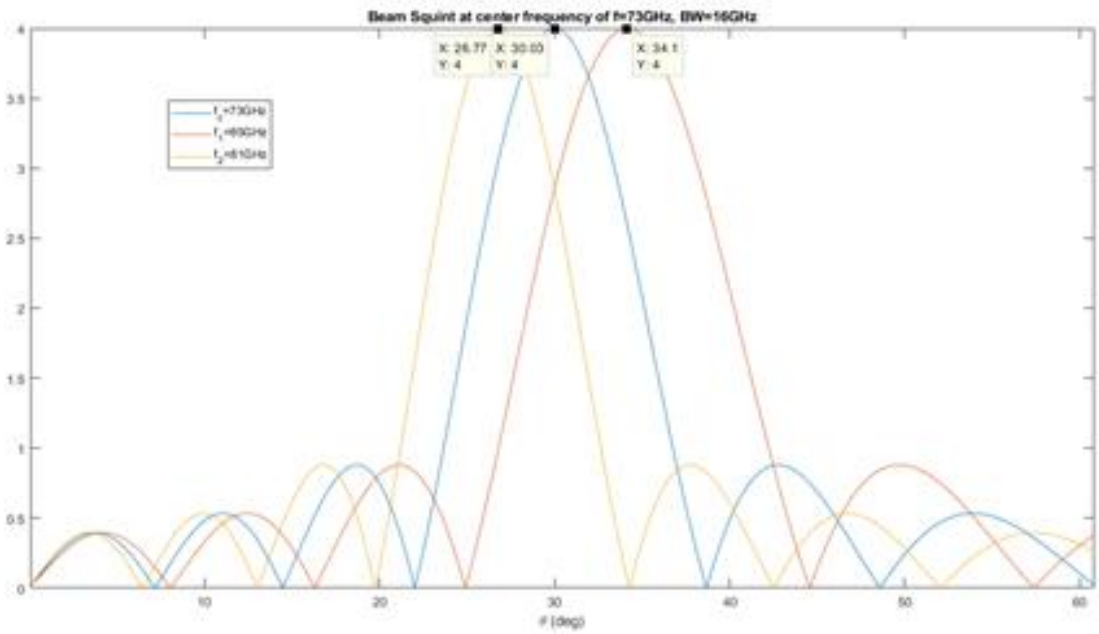


Figure 1: Beam squint in an array with 128 antenna elements, the antenna spacing is $d=\lambda_c/2$, where λ_c is the wavelength for the centre frequency.

[6] In a phased array or slotted waveguide antenna, squint refers to the angle that the transmission is offset from the normal of the plane of the antenna. In simple terms, it is the change in the beam direction as a function of operating frequency, polarization or orientation. It is an important phenomenon that can limit the bandwidth in phased array antenna systems.

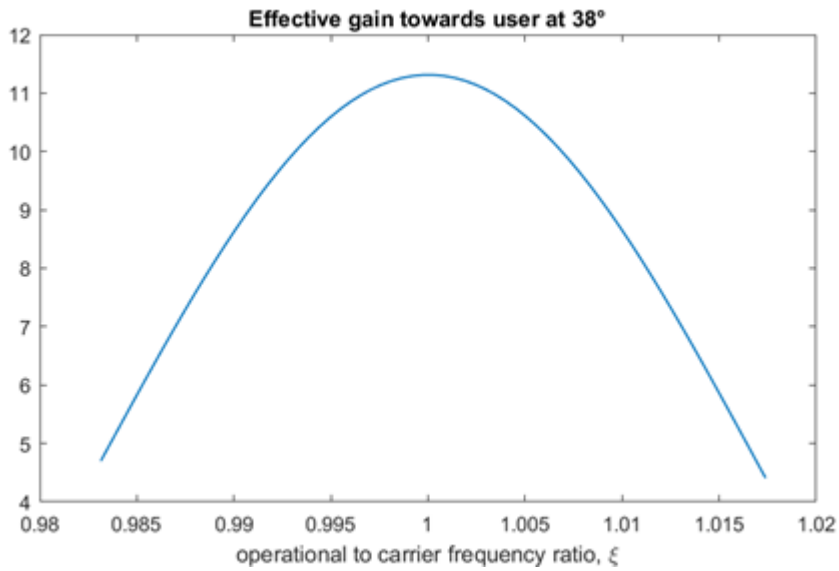


Figure 2: Channel frequency response for a system with 128 antenna elements, antenna spacing $\lambda_c/2$ and the expected beam direction at 38°.

Currently, most methods to deal with beam squint is to compensate it, based on beam squint estimation. While the estimation could be accurate, most of the methods are difficult to implement, due to the non-linearity of the time delay at different antenna elements. What is more, the existing best solution is a hardware implementation of true time delay (TTD) units [3]. These units are expensive to implement in mMIMO, as every antenna would require one. As such, a hardware implementation is infeasible.

Therefore, 5G-DRIVE investigated the beam squinting issue in a wideband orthogonal frequency division multiplexing (OFDM) system and proposed a resource allocation-based solution to compensate for the effect of beam squint. The principle of this method is as follows.

Conventionally, one user is only allocated with one beam to transmit one data stream. In the proposed scheme, multiple beams could be allocated to a user according to the beam squinting information collected. If the number of users is more than 1, a user is selected in a way that the centre beam points at it and the auxiliary beams can be squinted towards it. In a perfect scenario, all users would be spaced at max beam squint distance so as to reduce interference. Then, a data interleaving algorithm is applied to achieve a high system capacity. Figure 3 illustrates the performance of the proposed scheme for single and multi-user scenarios when the centre frequency is f_c = 73 GHz and system bandwidth is 2.5 GHz. The figure shows how the proposed allocation algorithm improves on a system that suffers from beam squinting. No beam squinting scenario (C_{NBS}) is included for comparison purposes. It shows the expected constantly rising capacity as the number of antennas increases. C_{BS} is a scenario in which the system suffers from beam squinting without any compensation. As the number of antennas increases, there is a rising trend before the system reaches a maximum capacity at ~120 antennas for a single user scenario and 64 antennas for a multi-user scenario when the number of users K is 3. As that number is passed, the capacity starts steadily decreasing. C_S is the proposed solution for the system. The performance is identical to C_{BS} until the maximum point, after which, (opposed to the beam squinting scenario) the trend continues increasing steadily. At the furthest point of analysis (256 antennas), there is an overall 80% increase in performance and the slope is approaching the curve, whereas the multiple user scenario experiences 150% performance increase. Figure 4 is provided for easier quantification of the performance increase. The reason multiple user performance increases by less than 300% of the single-user performance is due to the limitation imposed by the number of RF chains N_{RF}. A maximum of N_{RF}-2 users can be fully compensated on beam squint effects, the other users must reschedule some of their data to transmit it reliably.

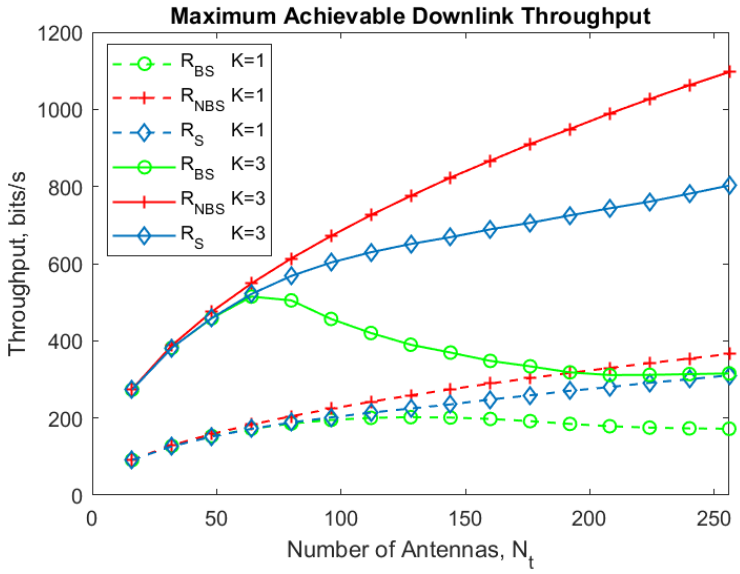


Figure 3: Maximum achievable capacity for an increasing number of antennas

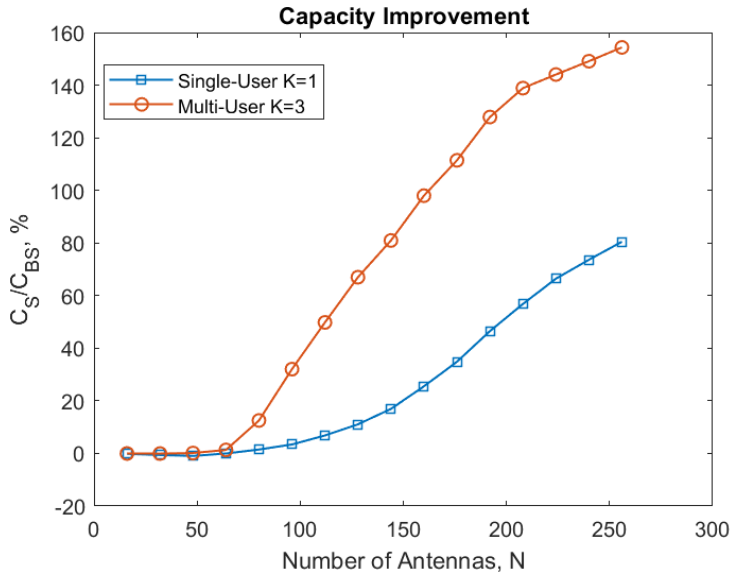


Figure 4: Capacity Improvement Multi-User *vs.* Single User. SNR for both curves is chosen to be 12dB, N_{RF}=3

Currently, a paper on resource allocation-based beam squint compensation has been submitted to an IEEE Conference[7].

2.2 Modelling of non-stationary channels for MIMO systems

In traditional channel modelling for MIMO systems, channels are considered to be stationary. However, according to the results obtained from trials, in MIMO systems, especially under the

[7] I. Laurinavicius, J. Wang, H. Zhu, and Y. Pan: "Beam squint exploitation for linear phased arrays in a mmWave multi-carrier system", submitted to IEEE Globecom 2019.

environment with mobility[8], the channel may not be stationary. Therefore, traditional channel model based on the assumption of stationary channel may not be suitable. In 5G-DRIVE, two approaches are considered to model non-stationary channels: (i) Dent model[9], also known as the modified Jake's model[10],[11], and; (ii) channel transformation into visual regions[12] (VRs) and non-visual regions (nVRs).

2.2.1 Applying the Dent Model with Conventional Water-Filling (CWF) Algorithm

In the Dent model (also known as Modified Jake's Model), it is assumed that M independent fading waveforms are required, each of which is composed of N sinusoids. With k-th fading waveforms denoted by $T_k(t)$, a novel model (Wu model) is proposed as the following:

$$T_k(t) = \sqrt{\frac{1}{N_0}} \sum_{n=0}^{N_0-1} \left(A_{k_1}(n) + kA_{k_2}(n)\right) [\cos(\omega_M cos\alpha_{n_k} \cdot t + \theta_n)] \qquad \text{Equation 1.}$$

where $k = 0,1,2,\cdots,M-1$. ω_M is the maximum Doppler frequency shift, $N_0 = \frac{N}{4}$, θ_n are the independent random phases, each of which is uniformly distributed in $[0,2\pi)$. A_{k_1} and A_{k_2} are the different orthogonal weighting functions, such as Walsh-Hadamard[13] sequences, $A_{k_1}(n)$ is the k_1-th orthogonal sequence in $n(+/-1)$, which satisfies:

$$\frac{1}{N_0} \sum_{n=0}^{N-1} A_{k_p}^*(n) A_{k_p}(n) = \begin{cases} 1, k = 1, p = q \\ 0. else \end{cases} \frac{1}{N_0}$$

$$k, l = 0,1,2,\cdots,M-1 \; p, q = 1,2 \qquad \text{Equation 2.}$$

The arrival angle α_{n_k} is similar to that of Li-Huang model, and can be represented as:

$$\alpha_{nk} = \frac{2\pi N}{N} + \frac{2\pi k}{MN} \qquad \text{Equation 3.}$$

The in-phase and quadrature components of any single fader have the same set of Doppler frequency shifts. The correlation independence between in-phase and quadrature components of any single fader is guaranteed by the orthogonal weighting functions. The uncorrelated properties of different fading waveforms are guaranteed by both the orthogonal weighting functions and different sets of Doppler frequency shifts.

---

8 L. Sanguinetti, A.L. Moustakas, E. Björnson, and M. Debbah (2015, March): Large System Analysis of the Energy Consumption Distribution in Multi-User MIMO Systems with Mobility. *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1730-1745.

9 See: P. Dent, G.E. Bottomley, and T. Croft (1993, June): Jakes Fading Model Revisited. *IEEE Electronics Letters*, vol. 29, no. 13, pp. 1162-1163.

10 Z. Wu (2004, July): Model of independent Rayleigh faders. *IEEE Electronics Letters*, vol. 40, no. 15, pp. 949-951.

11 Also see, *among others*: Y. LI, and Y.L. Guan (2000): "Modified Jakes' Model for Simulating Multiple Uncorrelated Fading Waveforms". In IEEE (Ed.), Proceedings of the 51st IEEE Vehicular Technology Conference (VTC2000 Spring), pp. 46-49. Tokyo, Japan, May 15-18, 2000.

12 A. Ali, E.D. Carvalho and R.W. Heath (2019, June): Linear Receivers in Non-Stationary Massive MIMO Channels with Visibility Regions. *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 885-888.

13 For more informative details also see, inter-alia: *https://en.wikipedia.org/wiki/Hadamard_code*

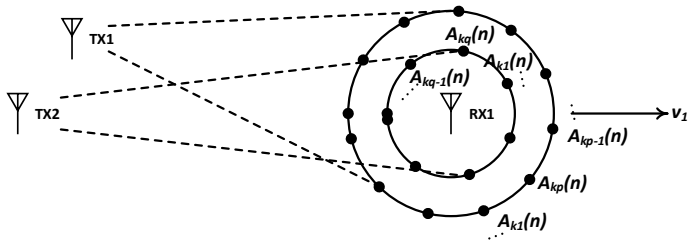


Figure 5: Dent channel model incorporating the Li-Huang model as M independent and orthogonal scatterers A_{kp} and A_{kq}, which are modeled by applying the Walsh-Hadamard sequences.

A simulation was run to generate two uncorrelated complex baseband fading waveforms, shown in Figure 5, using $N_0 = 8$ oscillators. At a carrier frequency of 28 GHz, a symbol rate of 64 ksps (kilo symbols per second) is chosen to obtain a fast fading channel at the receiver. The received signal strength indicator (RSSI) in dB is shown in Figure 6. A transmission symbol rate of 28 ksps is chosen to generate a fast fading channel. At a reviver velocity of 100 km/h, a Doppler spread of up to 150 kHz is observed, as shown in Figure 7.

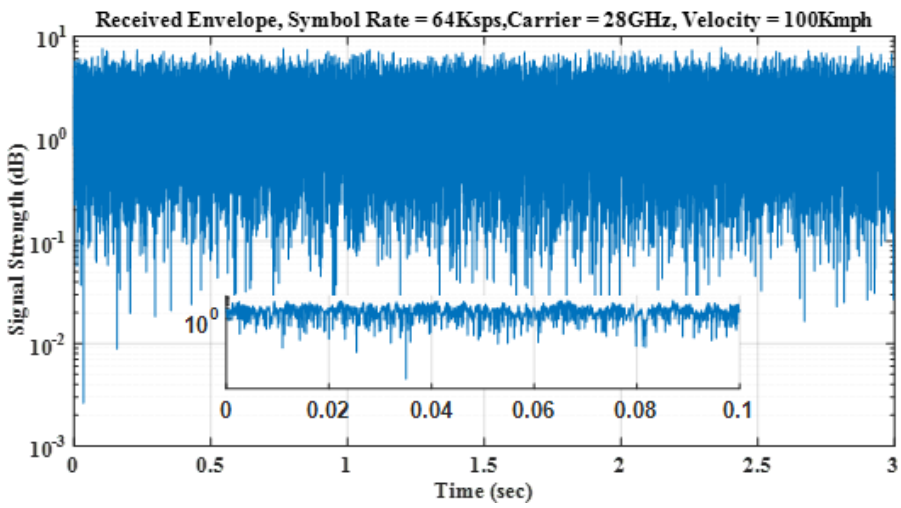


Figure 6: Received signal strength indicator (RSSI) of the Dent channel model at 28 GHz.

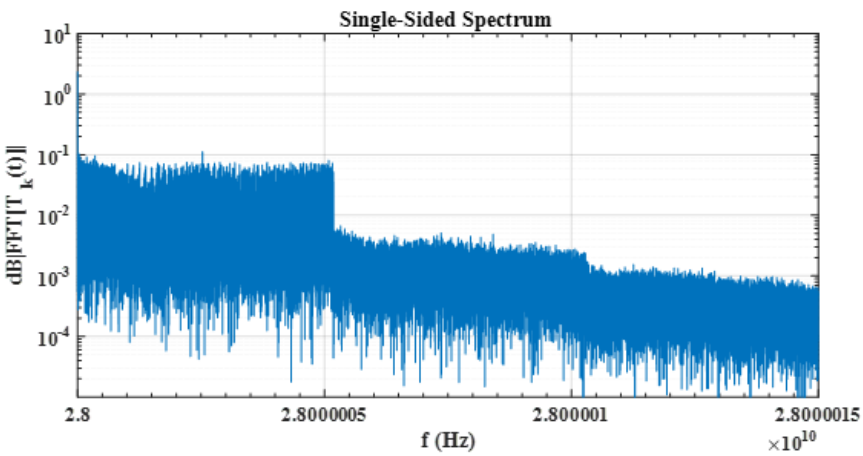


Figure 7: Single-sided spectrum (Doppler spread) of the Dent channel model at 28 GHz, indicating a spread of up to 150 kHz.

The quality of the fade's autocorrelation function can be measured by the mean-square-error defined by:

$$\epsilon^2 = \frac{\sum_{i=0}^{N-1}\left(R\left(\frac{i}{f_s}\right) - \frac{J_0\left(\frac{\omega_m i}{f_s}\right)}{2}\right)^2}{N} \qquad \text{Equation 4.}$$

where N denotes the time interval over which the mean-squared error is evaluated. Figure 8 plots the normalized autocorrelation of the first waveform at the receiver.

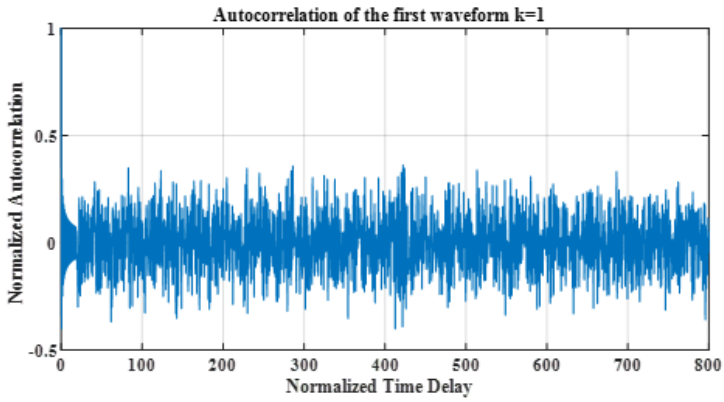


Figure 8: Normalized autocorrelation of the first waveform (k=1).

The quality of the cross-correlation function can also be measured by the mean-square-error defined by:

$$\epsilon^2 = \frac{\sum_{i=0}^{N-1}\left(R\left(\frac{i}{f_s}\right)\right)^2}{N} \qquad \text{Equation 5.}$$

Figure 9 shows the cross-correlation between the first ($k = 1$) and the second ($k = 2$) waveforms.

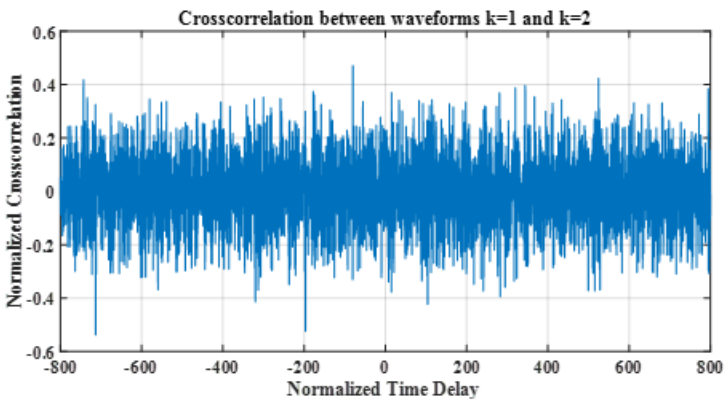


Figure 9: Normalized cross-correlation between the waveforms.

We propose quantizing the RSSI into two states: low-fade and deep-fade states by setting an RSSI threshold, and then applying a conventional water-filling (CWF) algorithm[14] at the receiver in order to accomplish equalisation. It is proposed that the receiver will apply equalization during the low fade state, with the receiver being idle during the deep fade state. Figure 10 shows the RSSI quantization threshold set to 6 dB, i.e. the receiver will apply CWF equalization only of the RSSI is at 6 dB or greater.

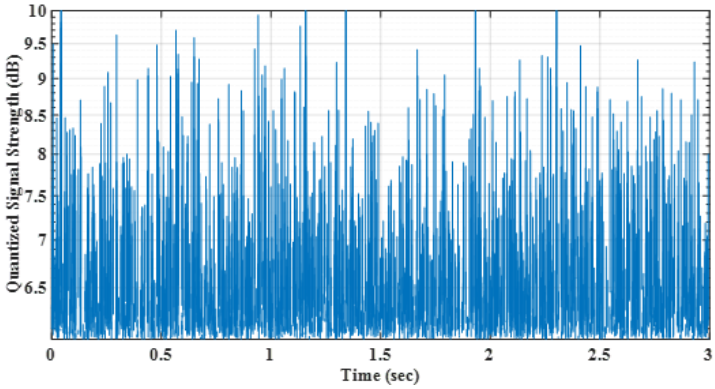


Figure 10: Quantized RSSI output at the receiver set to 6dB.

The spectral efficiency attained under CWF equalization at the receiver for a channel bandwidth B (=100 MHz, chosen for the simulation) is given by:

$$C = Blog_2\left(\frac{\gamma(t)}{\gamma_{th}}\right) \qquad \text{Equation 6.}$$

where $\gamma(t)$ is quantized the RSSI at time t and γ_{th} is the value of the RSSI threshold. Figure 11 plots the achieved spectral efficiency.

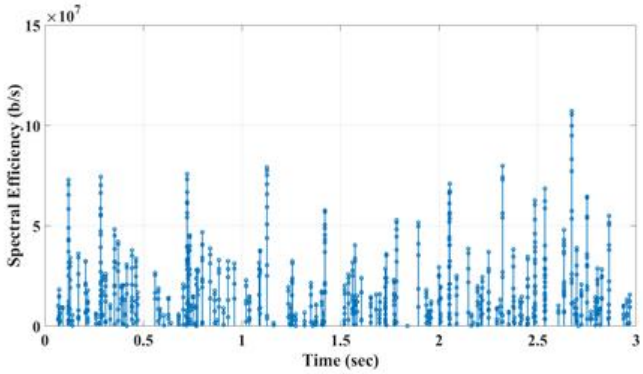


Figure 11: Spectral Efficiency with CWF equalization for a non-stationary channel.

[14] For further information also see: D.P Palomar, and J.R. Fonollosa (2005): Practical algorithms for a family of waterfilling solutions. *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 686-695.

2.2.2 Applying channel transformation into visual regions (VRs) and non-visual regions (nVRs) for non-stationary (NS) channels

The concept of VR/nVR is as follows: First, any channel model can be chosen (in our scenario, we will choose the nLoS cluster-ray distribution), $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_k, \cdots, \boldsymbol{h}_K]$. Then, the channel is "partitioned" into visual regions (VRs) and non-visual regions (nVRs) by introducing a matrix $\Xi_{\mathrm{k}} = \boldsymbol{D}_k^{1/2} \boldsymbol{R}_k \boldsymbol{D}_k^{1/2}$, where $\boldsymbol{R}_k$ is the channel autocorrelation matrix of user k, and $\boldsymbol{D}_k$ is a diagonal matrix such that if the signal transmitted from only D_k antennas (out of M antennas at the BS, $M > K$) is received at the user k, $\boldsymbol{D}_k$ has D_k non-zero diagonal entries. Then, if $\boldsymbol{z}_k \sim \mathcal{C}(\boldsymbol{0}, \frac{1}{M}\boldsymbol{I})$, $\boldsymbol{h}_k^{NS} = \sqrt{M}\Xi_k^{1/2}\boldsymbol{z}_k$ is the non-stationary (NS) channel at user k.

In the future, according to the test results obtained from WP3, this channel model will be explored further to retrieve the channel autocorrelation matrix and diagonal matrix under different test setup.

2.3 Deep learning for multi-user detection

Multiuser multiple-input multiple-output (MU-MIMO) signal detection over noisy fading channels is mathematically an integer least-squares (ILS) problem[15], which aims to minimize the pairwise Euclidean distance between the transmitted signal multiplied by channel matrix and the received signal. Concerning the maximum-likelihood solution to be computationally too expensive, it is usual practice to employ linear algorithms such as the matched filter (MF), zero forcing (ZF), and linear minimum meansquare error[16] (LMMSE) to trade off the optimality for lower computational complexity. Concerning the linear algorithms often too sub-optimum due to their uses of symbol-by-symbol detection, enormous research efforts have been paid in the last two decades, to achieve the best performance-complexity trade-off through the use of nonlinear sub-optimal algorithms such as V-BLAST[17], LMMSE-SIC[18], fixed-complexity sphere decoding[19] (SD), lattice-reduction (LR) aided detection[20] algorithms, and so forth. A quite comprehensive survey of the MIMO detection algorithms can be found in [4]; and we can reach two conclusions: (i) nonlinear algorithms are too complex to afford by the current DSP technology, and; (ii) they do not support well parallel computing which is however the trend of the future DSP technology for 5G and beyond. Recent advances towards the multiuser-MIMO detection problem lie in the use of deep learning.

In this subsection, unsupervised deep learning solutions for multiuser single-input multiple-output (MU-SIMO) coherent detection are studied. According to the ways of utilizing the channel state

---

[15] For more related information about the ILS also see, *inter-alia*: M. Al Borno (2011, February): *Reduction in Solving Some Integer Least-Squares Problems (Ph.D. Thesis)*. School of Computer Science, McGill University, Quebec, Canada. Available at: *https://arxiv.org/pdf/1101.0382.pdf*

[16] For further information about the LLMSE also see: *https://en.wikipedia.org/wiki/Minimum_mean_square_error*

[17] For more informative details about the V-BLAST algorithm also see, *among others*, the information presented in: *http://www.ee.columbia.edu/~jiantan/E6909/wolnianskyandfoschini.pdf*

[18] For more informative details about the LMMSE-SIC (Linear Minimum Mean Squared Error – Soft Interference Cancellation) algorithm also see, *among others*, the information presented in: Z. Luo, S. Liu, M. Zhao, and Y. Liu (2006): "A Novel Optimal Recursive MMSE-SIC Detection Algorithm for V-BLAST Systems". In IEEE (Ed.), Proceedings of the 2006 IEEE International Conference on Communications (ICC-2006). Istanbul, Turkey, June 12-15, 2006.

[19] For further informative details about the fixed- complexity sphere decoding algorithm also see, *for example*: L.G. Barbero, and J.S. Thompson (2008, June): Fixing the Complexity of the Sphere Decoder for MIMO Detection. *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2131-2142.

[20] For further informative details about the lattice-reduction aided detection algorithm also see, *inter-alia*: K.A. Singhal, T. Datta, and A. Chockalingam (2013): "Lattice reduction aided detection in large-MIMO systems". In IEEE (Ed.), Proceedings of the 2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC-2013). Darmstadt, Germany, June 16-19, 2013.

information at the receiver side (CSIR), deep learning solutions are divided into two categories. One is called equalisation-and-learning (EL), as depicted in Figure 12, which utilizes the CSIR (i.e. the MIMO channel matrix $\mathbf{H}$) for channel equalisation and then employ deep learning for multiuser detection (MUD). The other is called direct learning (DL), as depicted in Figure 13, which directly feeds the CSIR, together with the received signal, into deep neural networks[21] (DNN) to conduct the MUD [5].

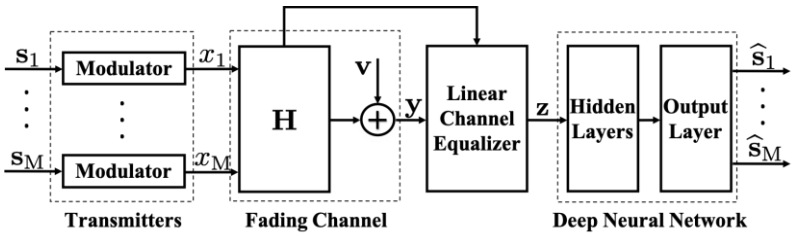


Figure 12: Block diagram of equalisation-and-learning for multiuser-SIMO detection.

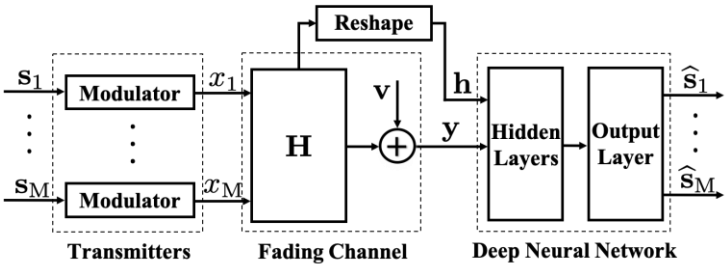


Figure 13: Block diagram of direct-learning for multiuser-SIMO detection.

Two experiments have been structured with respect to the EL mode and the DL mode. The performance is evaluated using the bit error rate (BER) averaging over sufficient Monte-Carlo[22] trials of block Rayleigh fading[23] channels and compared to widely used baselines under different channel conditions. The SNR is defined as the average received bit-energy to noise ratio (Eb=N0).

Experiment 1 (The EL Mode): In this experiment, we evaluate the BER performance of the CE-L mode under BPSK modulation[24]. Figure 14 illustrates the BER performance for the channel equalisation-

[21] For more informative details also see: Skymind Inc.: *A Beginner's Guide to Neural Networks and Deep Learning*. Accessible at: *https://skymind.ai/wiki/neural-network*

[22] Monte Carlo methods, or Monte Carlo experiments, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. The underlying concept is to use randomness to solve problems that might be deterministic in principle. They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other approaches. Monte Carlo methods are mainly used in three problem classes: optimization, numerical integration and generating draws from a probability distribution. For more informative details about the Monte-Carlo method also see, *for example*: *https://en.wikipedia.org/wiki/Monte_Carlo_method*

[23] Rayleigh fading is a statistical model for the effect of a propagation environment on a radio signal, such as that used by wireless devices. Rayleigh fading models assume that the magnitude of a signal that has passed through such a transmission medium (also called a communication channel) will vary randomly, or fade, according to a Rayleigh distribution - the radial component of the sum of two uncorrelated Gaussian random variables. For further information see, *inter-alia*: *https://en.wikipedia.org/wiki/Rayleigh_fading*

[24] BPSK (also sometimes called PRK, phase reversal keying, or 2PSK) is the simplest form of phase shift keying (PSK). It uses two phases which are separated by 180° and so can also be termed 2-PSK. It does not particularly matter exactly where the constellation points are positioned, and in this figure they are shown on the real axis, at 0° and 180°. Therefore, it handles the highest noise level or distortion before the demodulator reaches an incorrect decision. That makes it the most robust of all the PSKs. It is, however, only able to modulate at 1 bit/symbol (as seen in the figure) and so is

and-learning paradigm, where the channel equalizer can be linear ZF[25], LMMSE[26] or MF. The deep learning was conducted using mini-batch stochastic gradient descent[27] with Adam optimizer at the learning rate = 0.001. The loss function was categorical cross-entropy. The mini-batch size of each epoch was 1,000, and the deep neural network (DNN) was trained at Eb/No=5 dB; as this configuration is found to provide the best performance. The baseline for performance comparison is obtained using conventional symbol-by-symbol detection algorithm. It is shown that deep learning modules largely improve the performance of the MF-based receiver (around 8 dB at BER of 10E-3). The performance gain mainly comes from the sequence-detection. For both the linear-ZF and LMMSE receiver, the deep learning gain is almost negligible since there is no correlation among the equalised symbols; and thus, deep learning gain vanishes.

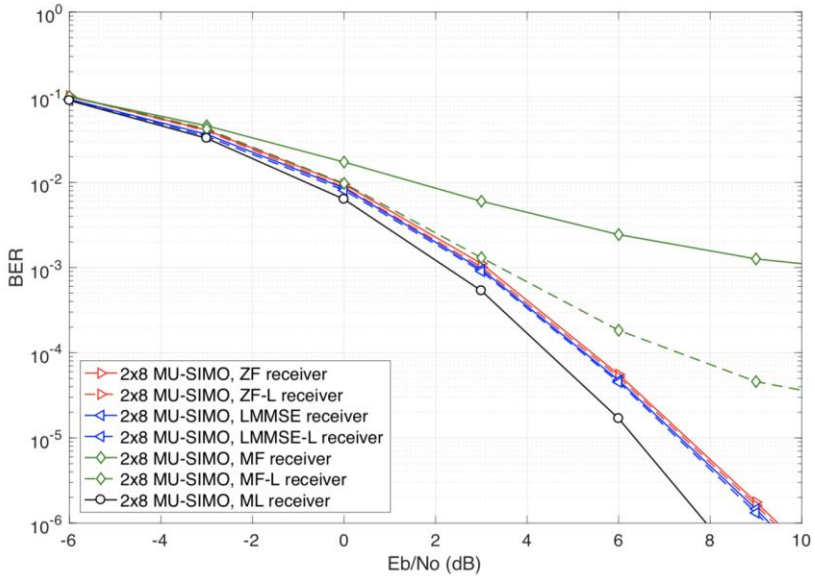


Figure 14: Average bit-error-rate performance of the equalisation-and-learning paradigm.

Experiment 2 (The DL Mode): In this experiment, we evaluate the BER performance of the DL mode under BPSK modulation. Figure 15 illustrates the BER performance of the Direct-Learning mode with different levels of CSIR. The baseline for performance comparison is the maximum likelihood receive. For the uncoded 2-by-8 multiuser-SIMO communications, it is found that DNN receiver fails to detect the transmitted waveform without CSIR since the received signals might have multiple combinations of the channel and transmitted signals; and such significantly degrades the deep-learning based classification performance. Consider the receiver having full CSIR; the Direct-Learning approach

unsuitable for high data-rate applications. For further details also see, *inter-alia*: *https://en.wikipedia.org/wiki/Phase-shift_keying#Binary_phase-shift_keying_(BPSK)*

25 Zero Forcing Equalizer refers to a form of linear equalization algorithm used in communication systems which applies the inverse of the frequency response of the channel. The Zero-Forcing Equalizer applies the inverse of the channel frequency response to the received signal, to restore the signal after the channel. It has many useful applications. For example, it is studied heavily for IEEE 802.11n (MIMO) where knowing the channel allows recovery of the two or more streams which will be received on top of each other on each antenna. The name Zero Forcing corresponds to bringing down the Intersymbol Interference (ISI) to zero in a noise free case. This will be useful when ISI is significant compared to noise.

26 For more informative details also see, *inter-alia*: D. Neumann, T Wiese, and W. Utschick (2017, July): Learning the MMSE Channel Estimator. *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2905-2917.

27 Stochastic gradient descent (often shortened to SGD), also known as incremental gradient descent, is an iterative method for optimizing a differentiable objective function, a stochastic approximation of gradient descent optimization. Also see, *for example: https://en.wikipedia.org/wiki/Stochastic_gradient_descent*

achieves near-optimal performance almost throughout the entire SNR range. The most interesting phenomenon is that the Direct-L approach can work with partial CSIR. Considering the CSIR to be known only for the user terminal (UT)1-to-receiver link, it is shown that UT1 enjoys a much better performance than UT2. The performance gap between UT1 and the maximum-likelihood receiver is caused by the co-channel interference from UT2. This result shows that deep learning utilizes the CSIR as the user signature to assist the signal classification.

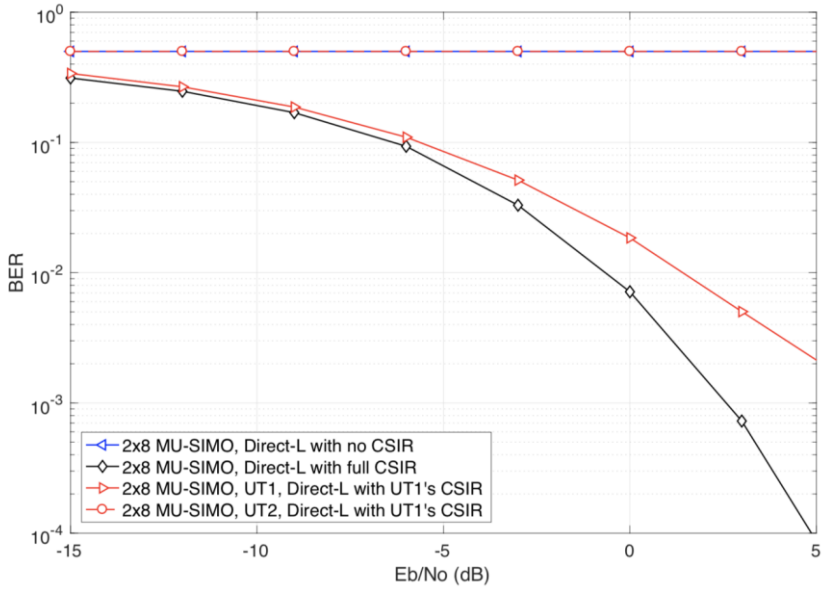


Figure 15: Average bit-error-rate performance of the directly-learning paradigm.

In addition to the above results, it is also found that the direct-learning solutions outperform the equalisation-and-learning solutions due to their better exploitation of the sequence-detection gain. On the other hand, the direct learning solutions are not scalable to the size of MU-SIMO networks, as current DNN architectures cannot efficiently handle many co-channel interferences[28]. This observation has motivated further research towards scalable direct-learning based multiuser-MIMO detection.

2.4 Hybrid user association under DU/CU architecture

In a 5G network with the distributed unit (DU) and central unit (CU), the user association can be done distributed with the DU or centralised at the CU that governs several DUs. If users can properly associate to the DU in the network, then the network can achieve the load balance, reduce the inter-cell inference and improve the capacity. The distributed approach has the advantage of scalability, but it may not provide the optimal solution for the whole network. The centralised approach can solve the optimization problem, however, with the cost of high overhead and thus low scalability. The hybrid solution can take advantage of both approaches. It uses the Central Control Unit (CCU) at CU to decide the association criteria for all DUs, and the DU to apply the criteria and make the association decision individually. The hybrid user association (HUA) problem[29] for load balancing of

[28] Also see: J. Wang, Z. Yi, J.M. Zurada, B.-L. Lu, and H. Yin (2006): *Advances in Neural Networks,* Springer.

[29] For further reading also see: H. Tian, W. Xie, X. Gan, and Y. Xu (2016): Hybrid user association for maximizing energy efficiency in heterogeneous networks with human-to-human/machine-to-machine coexistence. *IET Communications,* vol. 10, no. 9, pp. 1035-1043.

multitier cellular networks is studied. The proposed hierarchical HUA approach builds on a combination of decentralised user association (DUA) and auxiliary intervention of a central control unit (CCU). A major challenge with the CCU intervention is the time interval determined by a selected CCU control cycle during which the DUA must accept all users that satisfy the prevailing association criterion while proactively mitigating potential resource depletions. Consequently, the primary focus of this work is on relating the control cycle of the CCU intervention with the incipient resource depletions, according to a maximum allowed resource depletion probability. By uniquely combining a set of mathematical tools from stochastic geometry and queueing theory, a novel HUA method is proposed which evaluates the association bias values of the DUA according to a CCU-optimized load vector and enables tier-based resource depletion probability provisioning over finite control cycles. The trade-offs between the proposed HUA method and the standard DUA approach, are compared.

Following the spatiotemporal modelling approach from [6] [7], a region of a two-dimensional space L is considered where a set of user equipment (UE) is served by BSs B_k from the kth tier of a multitier network where k ∈(1, K). The kth tier BS units are randomly located following independent homogeneous Poisson point processes[30] (PPPs) Φ_k with an intensity ϕ_k determined by BS density in units/km^2. The 5G-DRIVE project focuses on the downlink of the multitier network where each BS of the kth tier transmits with power Pk. The total bandwidth $W = N_{rb}W_{rb}$ is divided into N_{rb} RBs of width W_{rb}. For a given realisation of Φ_k, new data flow requests from the UE set U arrive independently and fall within the space L and time window (t_0, t_{0+t}). The arrival process of the UE flow units follows a space-time PPP Θ with a spatiotemporal intensity determined by UE flow density in units/(s·km2). The temporal mean UE flow arrival rate of the network is $\lambda_{net} = \theta\mathcal{L}$. Let (k, i) be the pair denoting the i_{th} BS of the k_{th} tier used for short notation as (k, i)th BS and $(\cdot)_{ki}$ for subindexing. The results show that the UE arrival process after the UA in the (k, i)th BS is Poisson with intensity $\lambda_{ki} = p_{ki}\theta$ determined by units/s where p_{ki} is the scaling factor according to the effective coverage area of the (k, i)th BS. An assumption that the departure or service times of the UE flows from the (k, i)th BS are i.i.d random variables for which only the mean μ_{ki} and the variance σ^2_{ki} are needed to be known or estimated is made. The moments of the flow departure times depend on the prevailing flow file size and cell data rate in different tiers. The possible coupling between the UE departure times of the BSs can be incorporated via the BS-specific moments which corresponds to commonly used time-averaged interference method [8] [9].

The load state of a BS is modeled as the number of remaining RBs in the resource pool yet to be scheduled at time t_0+t. The number of available RBs in the (k, i)th BS can be represented as:

$$q_{ki}(t_0+t) = \min\{\max\{b_{ki}+x_{ki}(t_0+t)-y_{ki}(t_0+t), 0\}, B_{ki}\} \qquad \text{Equation 7}$$

where B_{ki} is the maximum number of RBs, $b_{ki} > 0$ is the initial number of remaining RBs at time t_0, and $b_{ki} \leq B_{ki} \leq N_{rb}$. The random processes $x_{ki}(t)$ and $y_{ki}(t)$ denote, respectively, the cumulative numbers of RB arrivals and RB departures in the (k, i)th BS. If $B_{ki} = N_{rb}$, the spectrum is fully reused by the BSs with the reuse factor of one. It is assumed that B_{ki} is a large number that complies with the current trend of modern cellular systems where the frequency ranges have rapidly increased. However, a large B_{ki} does not prevent congestion in hotspot scenarios where also a high number of users is requesting a service. If there are no UE associated in the (k, i)th BS, $q_{ki}(t) = B_{ki}$. On the other hand, if $q_{ki}(t) = 0$, a resource depletion event is declared.

It is obvious that the arrival process of the available RBs is determined by the departure process of

[30] In probability, statistics and related fields, a Poisson point process is a type of random mathematical object that consists of points randomly located on a mathematical space. For more details see: S.N. Chiu, D. Stoyan, W.S. Kendall, and J. Mecke (2013): *Stochastic Geometry and Its Applications*. John Wiley & Sons.

the set of UE from the cell whereas the RB departure process is determined by the UE arrival process. If the number of allocated RBs per UE is one for all users, there is one-to-one correspondence between an RB departure and a UE arrival. However, there can be M UE service classes where each class m∈(1, M) determines the required number of RBs per UE arrival Q_m, according to a selected bandwidth allocation strategy summarized. The mean and variance of the requested number of RBs per arriving UE (i.e. a batch arrival of RBs) are denoted as ϑ_{ki} and ν_{ki}, respectively.

2.4.1 RDP analysis at a BS level

The following presents the resource depletion probability (RDP) analysis methods in different hierarchy levels (i.e. BS, tier, and network) and relate the user arrival rate with the maximum allowable RDP and finite horizon. The results are then applied to the HUA approach. In essence, the RDP acts as a statistical quality of service parameter to indicate if the service is going to be temporally blocked due to lack of RBs in the BSs. It is extremely difficult to obtain exact analytical approaches for the finite-horizon RDP with the RB model presented above. Consequently, the diffusion approximation approach is applied, which has been successfully used to find a closed form solution to many multidisciplinary queueing theoretic problems [10]. In this approach, a discrete queue model is replaced by a continuous time Brownian model[31] that is characterized by the first- and second-order moments of the arrival and departure processes of the queue. The RDP over a finite time horizon T of the (k, i)th BS can be represented as in [11].

$$\begin{aligned}\mathcal{P}_{ki} &= \Pr(\tau_{ki} < T) \\ &= \int_0^T \frac{b_{ki}}{\sqrt{2\pi\alpha_{ki}t^3}} \exp\left[-\frac{(\beta_{ki}t + b_{ki})^2}{2\alpha_{ki}t}\right] \mathrm{d}t \\ &\leq \exp\left(-\frac{2\beta_{ki}b_{ki}T + b_{ki}^2}{2\alpha_{ki}T}\right)\end{aligned}$$

Equation 8

where τ_{ki} = inf(t≥ 0 | $q_{ki}(t) = 0$; $q_{ki}(0) = b_{ki}$) is the first passage time of $q_{ki}(0) > 0$ to $q_{ki}(t) = 0$, inf(·) is the infimum function, and β_{ki} and α_{ki} are, respectively, the mean and variance of the applied RB diffusion process.

2.4.2 Proposed HUA approach

The following approach presents a novel HUA method that dynamically evaluates the association bias values of the fixed DUA approach according to a CCU-optimized load vector and enables tier-based RDP provisioning over finite control cycles.

A. DUA

We first outline the standard DUA approach, based on the maximum average received power (RP) and fixed preassigned biasing, also known as the cell range extension[32] method. Specifically, in the DUA approach, the j_{th} UE is associated with the (k, i)th BS that satisfies:

---

[31] In mathematics, the Wiener process is a continuous-time stochastic process named in honor of Norbert Wiener. It is often called standard Brownian motion process or Brownian motion due to its historical connection with the physical process known as Brownian movement or Brownian motion originally observed by Robert Brown. For further information also see, inter-alia: *https://en.wikipedia.org/wiki/Wiener_process*

[32] A major issue in HetNet planning is to ensure that the small cells actually serve enough users. One way to do that is to increase the area served by the small cell, which can be done through the use of a positive cell selection offset to the received power of the Small Cell (SC). Extending the coverage of a cell by means of connecting a UE to cell that is weaker than the strongest detected cell is referred to as cell range extension.

$$\max_{k,i} \epsilon_k P^{\text{rx}}_{kij}$$

where ϵ_k is the fixed bias factor for the kth tier BSs and P^{rx}_{kij} is the average RP at the jth UE received from the $(k, i)^{th}$ BS which can be measured using a pilot signal transmitted by the BS. The unbiased DUA approach, where $\epsilon_k = 1$ in (c), is the optimal UA for maximizing the SINR coverage if all BSs are transmitting. For light traffic load scenarios, there is also a marginal improvement on SINR coverage via biasing [12]. However, the primary objective of the biasing is to balance the network load so that a UE is not blocked due to lack of RBs in the BSs.

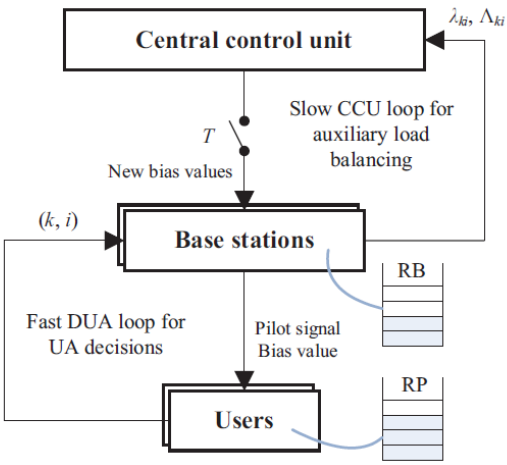


Figure 16: Proposed HUA approach.

B. HUA

The basic idea of the HUA approach is to split the UA process into two control loops, namely the slow CCU loop for updating the bias values ϵ_k in (c) and the fast DUA loop to perform the final UA decisions, see Figure 16. The primary focus of the proposed HUA is on relating the expected cycle T of the CCU intervention with the maximum allowable RDP. The CCU control cycle T, which is decided and broadcasted by the CCU, is the time that the DUA must wait until the new bias values are activated. The proper selection of T is affected by the expected changes in the statistical system parameters presented before that affect the RB depletion. By ensuring a proper minimum value for the CCU control cycle T, the ping-pong problem [13] can be mitigated. On the other hand, to avoid unnecessary bias updates and to effectively adapt the CCU control cycle, the CCU can employ a simple incipient congestion indicator as $I_{ki} = 1$, if $\lambda_{ki} > \Lambda_{ki}$ and $I_{ki} = 0$, otherwise. Specifically, if $I_{ki} = 0$, a bias update is not needed for the prevailing control cycle. It is also obvious that if $I_{ki} = 1$, a bias update cannot help significantly as the whole network is congested. The DUA decision loop cycle, on the other hand, is proportional to the mean UE inter-arrival time λ^{-1}_{net}. It is a reasonable assumption that $T \gg \lambda^{-1}_{net}$. The association bias values ϵ_k in the CCU are found in two steps. The first step evaluates the UE arrival rates per cell that satisfy the arrival rate constraint while leading to a minimum number of offloading with respect to the unbiased DUA. In other words, the step directs a UE to access a BS with maximum RP when there are enough RBs left to support a maximum allowed RDP. In the second step, the bias factors are found using the arrival rates from the first step. A pseudocode of the HUA approach is given in Figure 17.

```
Initialization: ϵ_k = 1, ∀k
for each CCU control cycle with duration T do
    Observe λ_ki and Λ_ki in each BS and inform the CCU
    if I_ki = 1 for any k, i then
        Trigger the CCU to obtain new bias values
        Update λ_net = Σ_{k,i} λ_ki
        Calculate new arrival rates λ* using STEP 1
        Calculate new bias values ϵ* using STEP 2
        Deliver new bias values to BSs
    end if
    Perform the DUA with the updated biases using (10)
end for
```

Figure 17: Pseudocode of the HUA approach.

C. Results

The following numerical examples focus on a two-tier network with K=2 in order to obtain presentable plots of tier-specific results for different UA methods. All connection requests are associated without any UE blocking prior potential resource depletions. The selected network area is $\mathcal{L} = \pi r^2$ with radius r = 1000 m. The tier-based transmission powers of the BSs are set to $\{P_1, P_2\} = \{53, 33\}$ dBm. It is assumed that there are N_{rb} = 400 RBs to be allocated with W_{rb} = 180 kHz and cell frequency reuse factor of one. The SINR outage threshold is set to 0 dB and Rayleigh fading models the random fluctuation of the channel. In order to demonstrate the inherent load balancing effects, different initial load conditions are assigned for each tier as $\{b_{1i}, b_{2i}\} = \{400, 100\}$. The tier-dependent UE inter-departure times are assumed to be exponential with rates $\{\mu_{1i}, \mu_{2i}\} = \{0.1, 0.4\}$, and variance $\{\sigma^2_{1i}, \sigma^2_{2i}\} = \{100, 6.25\}$. To demonstrate the random RB batch process, three UE service classes are used, each class requiring $\{Q_1, Q_2, Q_3\} = \{1, 5, 10\}$ RBs per flow arrival.

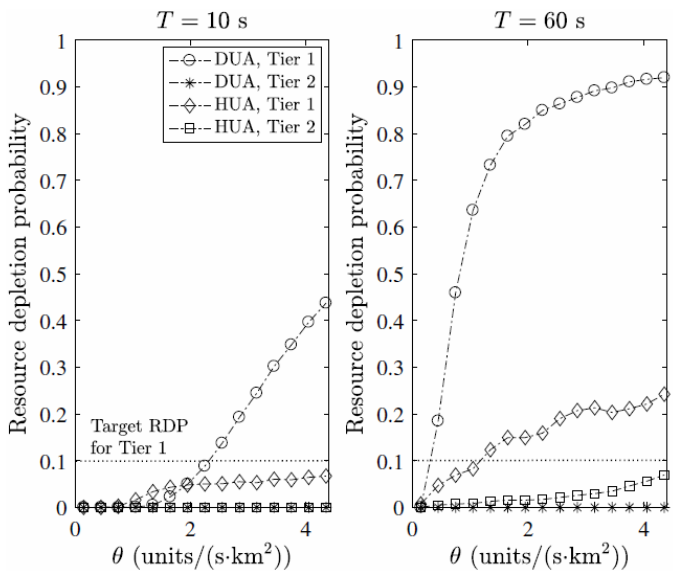


Figure 18: Comparison of RDP for different UA methods.

Figure 18 compares the simulated tier-level RDPs of the DUA and HUA as a function of θ with two different control cycles T=10 s and T=60. For the DUA, typical fixed bias factors $\{\epsilon_1, \epsilon_2\} = \{0,10\}$ dB are used, setting $\{\phi_1, \phi_2\} = \{1/\mathcal{L}, 10/\mathcal{L}\}$. The target RDP per tier is set to be $\{\rho_{1i}, \rho_{2i}\} = \{0.1, 0.01\}$. It is seen that the DUA clearly exceeds the target maximum RDP by allowing too many UE to associate with the tier-1 BSs while the HUA approach better supports the RDP. The demonstration shows that load balancing is effective if the network is unbalanced but not fully congested and there is a

sufficiently short CCU cycle (cf. the case with T = 10 s). If the network is fully congested or the CCU cycle is too long (cf. the case with T = 60 s), not all UE can be served at the time and their connection should be blocked in order to support the target RDP for the remaining users.

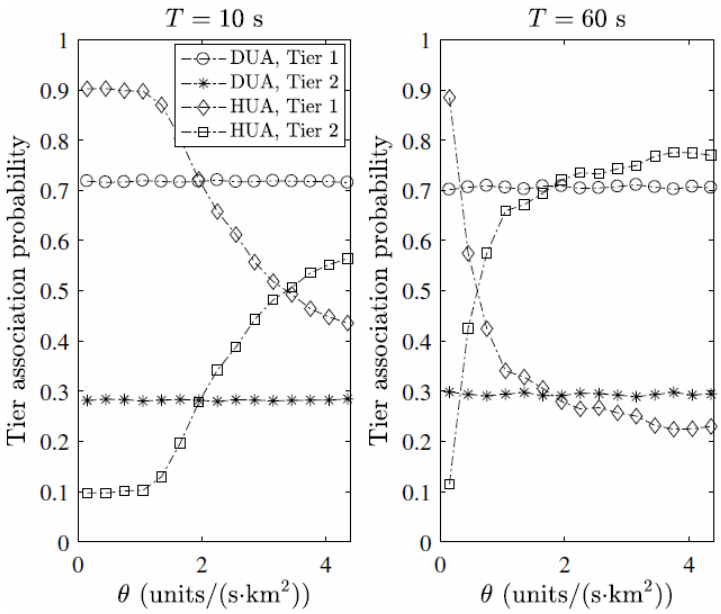


Figure 19: Comparison of tier association probability for different UA methods.

The probability of UE association is then evaluated in Figure 19 to further reveal the differences between the DUA with the preassigned biases and HUA with the RDP-aware biases. It is seen that as the value of θ is increased, the association probability per tier remains constant for the DUA whereas the HUA can modify the association probability between tiers to avoid exceeding the target RDP. Finally, the probability of SINR outage is studied in Figure 20. It is seen that the biasing increases the SINR outage probability, as the users are no longer allowed to associate with the BS offering the maximum received power. However, as seen from Figure 20, the HUA approach can reduce the SINR outages compared to the DUA by avoiding unnecessary offloading in the case the given RDP is satisfied.

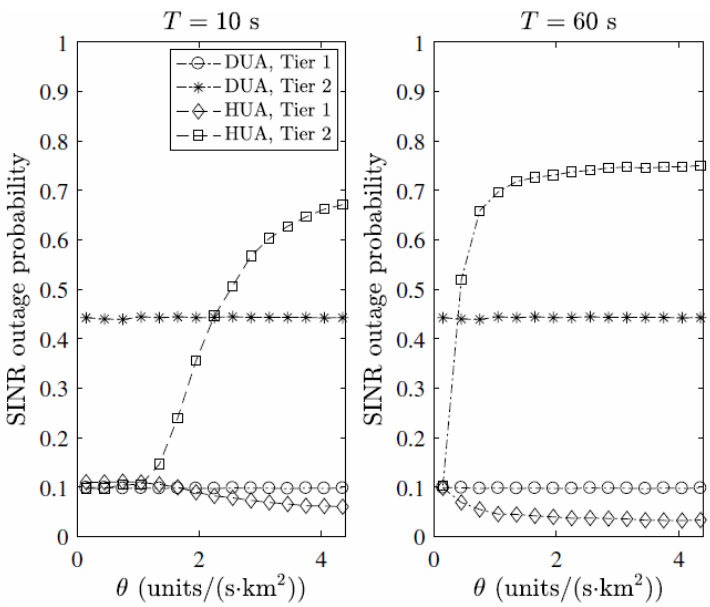


Figure 20: Comparison of SINR outage probability for different UA methods.

3 Transport networks

The requirements for the Transport Network in 5G are changing, due to the different interfaces now specified. Figure 21 shows the multitude of different interfaces specified by 3GPP[33] taking into account different core networks (4G or 5G), 4G or 5G base station (eNB[34], gNB[35]) connection to the possible core networks, control plane and user plane splitting (CUPS) which is now expected to extend into the Radio Access Network (RAN), and the RAN functional splitting of the base station into central unit (CU), distributed unit (DU) and remote radio unit (RRU) [14].

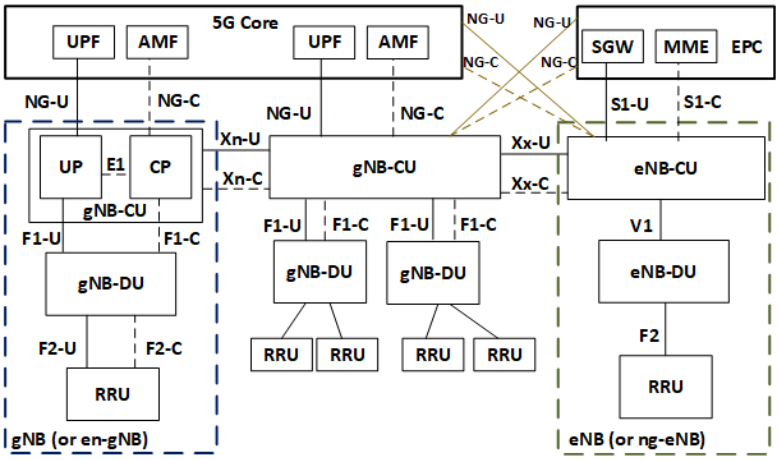


Figure 21: New transport interfaces defined by 3GPP for 5G networks.

In Figure 22, a visualisation of the transport for the 5G network with the different RAN functional split interfaces is presented. A virtualised network enabling dynamic resource provision and end-to-end slicing, key targets of 5G, is illustrated. Backhaul is provided from the virtualised CUs to the 5G and/or 4G (vEPC) core networks. A "middlehaul" form of xhaul[36] connects the vCUs with vDUs, using a higher-layer split (HLS) of the RAN functions[37] - the F1 interface[38] shown in Figure 21. In the case of

[33] The 3rd Generation Partnership Project (3GPP) unites seven telecommunications standard development organizations (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC), known as "Organizational Partners" and provides their members with a stable environment to produce the Reports and Specifications that define 3GPP technologies. The project covers cellular telecommunications network technologies, including radio access, the core transport network and service capabilities - also including work on codecs, security, quality of service (QoS)- and thus provides complete system specifications. The specifications also provide hooks for non-radio access to the core network, and for interworking with Wi-Fi networks. For more informative details also see: *https://www.3gpp. org/about-3gpp*

[34] E-UTRAN Node B, also known as Evolved Node B (abbreviated as eNodeB or eNB), is the element in E-UTRA of LTE that is the evolution of the element NodeB in UTRA of UMTS. It is the hardware that is connected to the mobile phone network that communicates directly wirelessly with mobile handsets (UEs), like a base transceiver station (BTS) in GSM networks. Traditionally, a Node B has minimum functionality, and is controlled by a Radio Network Controller (RNC). However, with an eNB, there is no separate controller element. This simplifies the architecture and allows lower response times.

[35] The next generation NodeB (gNB) is a concept originating for the 5G access network as assessed on 3GPP Release 15 and onwards. This node provides New Radio (NR) user plane and control plane protocol terminations towards the UEs and it is connected via the next generation (NG) interface to the 5G Core (5GC) network. For further information also see: *https://www.3gpp. org/release-15*

[36] For further information also see, *inter-alia*: A. de la Oliva, X, Costa-Pérez, et *al.* (2015): Xhaul: Toward an integrated fronthaul/backhaul architecture in 5G networks. *IEEE Wireless Communications,* vol. 22, no. 5, pp. 32-40.

[37] For further information also see: R. McKenzie (2018): *NGMN Overview of 5G RAN Functional Decomposition*. NGMN Alliance.

[38] For more details also see, *inter-alia*: The Third Generation Partnership Project (3GPP): TS 38.470 V15.4.0 (2018-12): "NG-RAN – F1 general aspects and principles".

Figure 22, Ethernet transport and aggregation are imagined for the "middlehaul". More detail on the HLS and lower layer split (LLS) is provided in the following section. The HLS/F1 interface is also used for "dual-connectivity"; it is the interface used to connect micro-base stations to macro-base station and this is shown by the midhaul transport. The term "fronthaul" is now used for the transport segment between the vDUs and the RRUs, or RAUs (Remote Antenna Units) in Figure 22, considered in a more generic sense than their definition by 3GPP. The LLS/F1 interface may be used here, and again may use Ethernet transport and aggregation, but analogue techniques may be used to reduce the bandwidth requirements on the fronthaul links. Note also the possibility of moving multi-access edge computing (MEC) facilities nearer to the user, at the vCUs, or in aggregation nodes up to the vDUs, reducing reliance on cloud computing in the core.

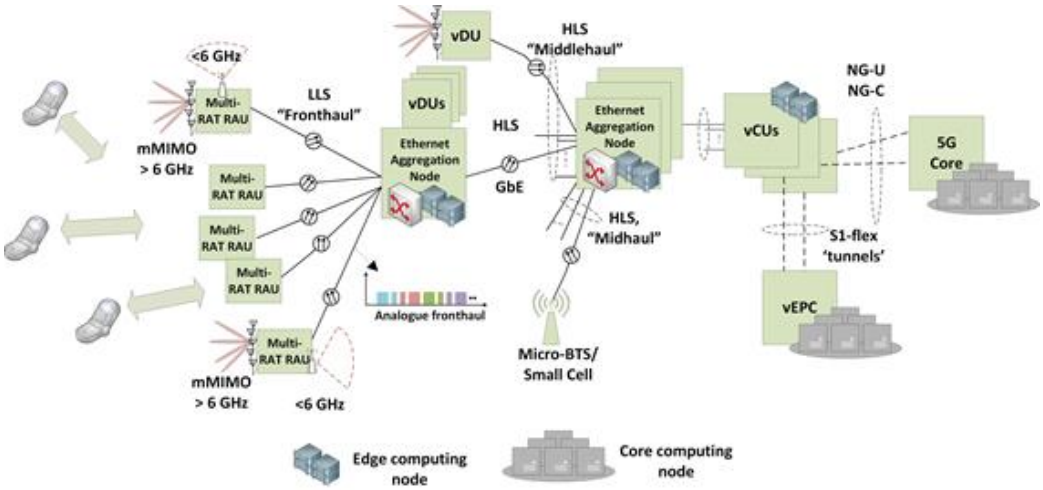


Figure 22: Visualisation of 5G transport network.

3.1 Split-RAN segments and requirements

This section provides further background information on the definition of the RAN functional splits, and the requirements on the transport network for these split points.

The 3GPP has defined the HLS or F1 interface as being between the PDCP (Packet Data Convergence Protocol) and RLC (Radio Link layer Control) functions of the Radio Access Network protocol stack. This is shown in Figure 23. This split point was previously defined as the Option 2 Split[39] by 3GPP when a whole range of different split points was being considered. Note that there is a clear CUPS at this split point, and this leads to multiple deployment options with a mix of centralised and distributed control plane and user plane processing.

[39] Also see: The Third Generation Partnership Project (3GPP): Technical Specification Group Radio Access Network; TR 38.806 V15.0.0 (2017-12): "Study of separation of NR Control Plane (CP) and User Plane (UP) for split option 2; (Release 15)".

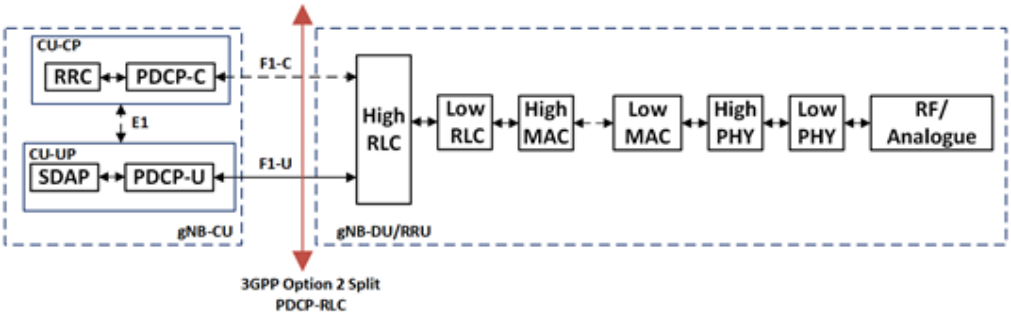


Figure 23: HLS/F1 interface in 3GPP RAN protocol stack.

The 3GPP was unable to decide on a specific LLS. Other organisations such as Open RAN[40] have continued the work of examining this split and even consider variable split points. Figure 24 shows possible options, by using the original 3GPP numbering for split point options, for the LLS or F2 interface. Option 8 is the "traditional" interface, used by CPRI[41], OBSAI[42], which requires the transport of time-domain samples of the radio waveform(s). Generally, as the split point is moved towards the left of Figure 24, there is a loss of centralisation (and potentially for function virtualisation), and an increasing difficulty to carry out joint processing of signals for MIMO and distributed MIMO (Coordinated Multipoint Joint Transmission/Joint Reception). However, latency (and latency variation) constraints on the fronthaul may reduce. Of greatest significance for these split points is the bit-rate reduction, first in transporting frequency domain rather than time domain samples (Option 7.1) and then by transporting the flows only for separate spatial streams (layers) (Option 7.2) and for separate users (Option 7.3, Option 6).

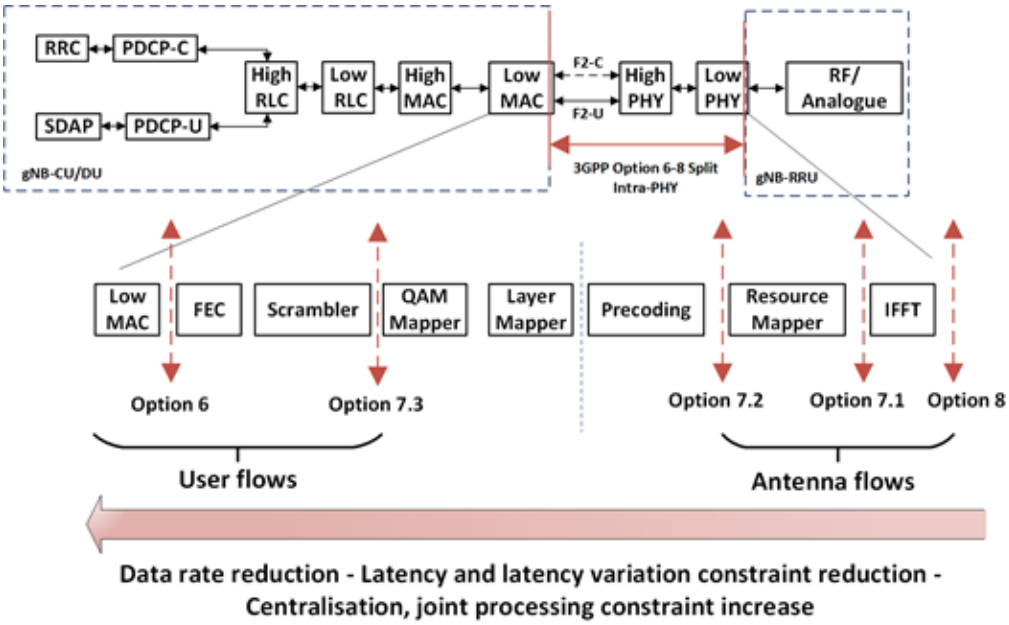


Figure 24: Possible LLS/F2 interfaces for 5G networks.

[40] Open RAN, Telecom Infra Project, *https://openran.telecominfraproject.com/*; see also ORAN Alliance, *https://www.o-ran.org/*. [Accessed April 2019].

[41] The Common Public Radio Interface (CPRI) standard defines a flexible interface between Radio Equipment Controllers (REC) and Radio Equipment (RE). Also see, *among others*: *http://www.cpri.info/spec.html*

[42] For more details about the Open Base Station Architecture Initiative see: *http://www.obsai.com/*

Further detail on the data rate, latency and other requirements on the different split points is provided in Table 1 [15]. The bit-rate requirements generally increase with greater centralisation, through both overheads and the need to send multiple spatial flows for MIMO. For options 7.1 and 8, there is also no dependency on cell load, and therefore little opportunity for statistical multiplexing gains. The one-way latency for the LLS/F2 is largely dependent on HARQ[43], which is variable for 5G, CSI[44] lifetime and radio synchronisation requirements. For the HLS/F1, it is mainly application dependent. Also shown in Table 1 are the requirements for an analogue fronthaul. Such systems have been used in legacy mobile networks mainly by "neutral host providers", through distributed antenna systems (DASs) for indoor or large venue coverage for multiple mobile operators. Analogue transport is of interest for 5G because of the bandwidth requirements in transporting massive MIMO signals.

Table 1: *Comparison between RAN split point requirements, adapted from* [15].

<table>
<tr><th>Split option</th><th>2</th><th>6</th><th>7.3</th><th>7.2</th><th>7.1</th><th>8</th><th>Analogue</th></tr>
<tr><td>Timing</td><td>NRT</td><td colspan="5">RT</td></tr>
<tr><td>One-way Latency</td><td><5 ms</td><td colspan="5">50-200 µs</td></tr>
<tr><td>Cell coordination</td><td>Distributed</td><td colspan="5">Centralised</td></tr>
<tr><td>HARQ</td><td colspan="2">Distributed</td><td colspan="4">Centralised</td></tr>
<tr><td>CUPS</td><td>YES</td><td colspan="3">TBD</td><td colspan="2">NO</td></tr>
<tr><td>Latency variation constraints</td><td>Relaxed</td><td colspan="3">Implementation dependent (constraint by one-way latency)</td><td>Tight (~10 ns)</td><td>Tight</td></tr>
<tr><td>Comparative data rate requirements</td><td>Low</td><td colspan="5">Increasing
----
-></td><td>Highest spectral efficiency</td></tr>
<tr><td>Data rate (Gbps)[3]</td><td>10-50</td><td colspan="5">+ 20-40% overhead ||</td><td>55 GHz[45]
4 GHz[46]</td></tr>
<tr><td>Flows</td><td colspan="2">User</td><td colspan="5">Antenna</td></tr>
<tr><td>Data rate dependency</td><td colspan="2">Cell load</td><td colspan="5">Cell load (not for 7.1, 8)/
num. antenna</td></tr>
</table>

3.2 Evolved/Ethernet fronthaul

A MAC-PHY functional split for an LTE RAN has been implemented and tested over an Ethernet fronthaul, see Figure 25 [16]. Figure 25 shows the functional decomposition carried out, performed

[43] Hybrid Automatic Repeat Request (hybrid ARQ or HARQ) is a combination of high-rate forward error-correcting coding and ARQ error-control. For more informative details and other related aspects also see, *among others*: *https://en.wikipedia.org/wiki/Hybrid_automatic_repeat_request*

[44] In wireless communications, Channel State Information (CSI) refers to known channel properties of a communication link. This information describes how a signal propagates from the transmitter to the receiver and represents the combined effect of, for example, scattering, fading and power decay with distance. The method is called as channel estimation. The CSI makes it possible to adapt transmissions to current channel conditions, which is crucial for achieving reliable communication with high data rates in multi-antenna systems.

[45] For the case of 128T/128R MIMO.

[46] For the case of analogue beamforming.

using Open Air Interface[47] (OAI) software on a PC for the RAN functions. At this split point, control plane signals (what would be primitive exchanges in closed operation of the RAN functions) are clearly visible: examples shown are the System Information (SI), Downlink Control Information (DCI) and Random-Access Request (RAR) messages (there are others), while the Downlink Shared Channel (DLSCH) is indicated as carrying the user data. In the Ethernet fronthaul implemented, the different packets can be identified, queued differently and sniffed as they are transported.

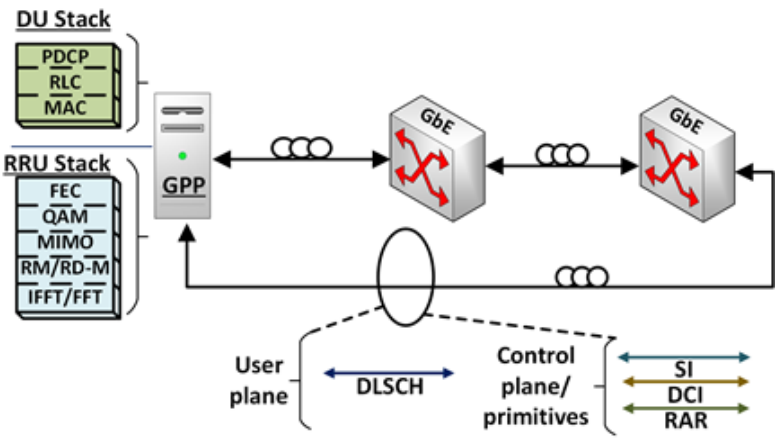


Figure 25: MAC-PHY (Option 6) Lower-layer Split Implementation with Ethernet fronthaul.

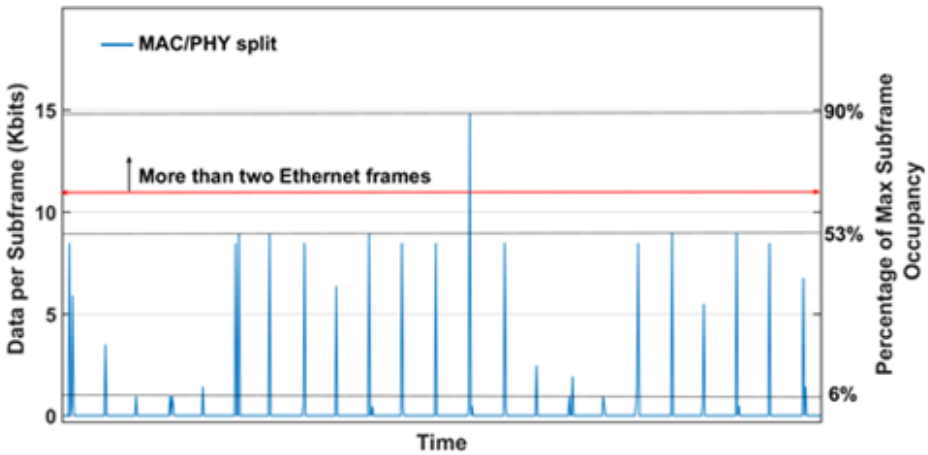


Figure 26: MAC-PHY (Option 6) Lower-layer Split throughput measurement.

Within the Ethernet fronthaul, measurements can be taken of data transfer/throughput and latency, the latter using in-line hardware probes for greater accuracy [17]. Figure 26 shows the user data transfer in time per LTE subframe for the MAC-PHY split, indicating individual frame transmissions, and instances where there more than one Ethernet frame needs to be sent in close succession. The variability in the LTE subframe size, and the number of Ethernet frames required to transmit the data is due to the dependence on the cell load at this split point. A per-flow mapping then allows statistical multiplexing gains to be achievable.

[47] For more details also see: *https://www.openairinterface.org/*

Table 2: Latency measurements for MAC-PHY split Ethernet fronthaul.

KPI	1 GbE	10 GbE
Average packet latency/ µs	45	43
Packet latency STD/ µs	9	6.9
Average subframe latency/ µs	54	52
Subframe latency STD/ µs	6.3	4.1

Table 2 shows latency measurements for the Ethernet fronthaul, in terms of average latency and its standard deviation[48] (STD). This is shown for the Ethernet frames and for the LTE subframes that they transport. Although, the LTE subframe latency is generally increased as some Ethernet frames that are needed for constructing the LTE subframe take longer than others to arrive, the standard deviation generally reduces as the statistical variations in the arrivals of the Ethernet frames are averaged out. A comparison is also carried out for 1 Gb Ethernet and 10 Gb Ethernet transport, showing the reduction in latency and latency variation for the latter. Future work will examine the use of Time-Sensitive Networking[49] (TSN) techniques to prioritise timing and synchronisation packets, and primitives (thus extending the CUPS into the lower layers and requiring separate mapping of these into the Ethernet packets). Although, the performance in Table 2 conforms generally to requirements, the measurements were performed in a lightly loaded/aggregated network [17]. Large parts of the latency may also be due to software implementations, and hardware offloading will be performed for key RAN functions.

3.3 Analogue fronthaul

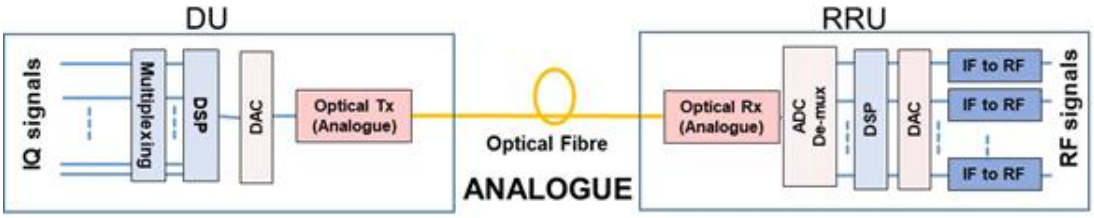


Figure 27: DSP-assisted analogue fronthaul.

As stated previously in Section 3.1, analogue transport for the final part of the fronthaul (e.g. F2/LLS or even a further final drop to the antennas) is now of interest, since it is highly spectrally efficient. One problem with traditional analogue subcarrier multiplexing[50] (SCM) techniques in fibre optic communications has been the reliance on microwave components (mixers, filters, etc.) which leads to a lack of flexibility and reconfigurability. The use of digital processing in DU and RRU (assuming a F2 split) enables adaptation/scalability to variable signal bandwidths and multiplex sizes, simple incorporation of different modulation techniques (single- and double-sideband), and the possibility

[48] In statistics, the standard deviation (STD) is a measure that is used to quantify the amount of variation or dispersion of a set of data value. A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

[49] For more details about TSN techniques also see: *https://www.tttech.com/technologies/time-sensitive-networking-tsn/?gclid=Cj0KCQjw19DlBRCSARIsAOnfRehgTjAgvYD5rP7IJ_la5ptocIuKT6vymD9s7FIN1DvXIOEpO9ycuQIaAgOREALw_wcB*

[50] Subcarrier Multiplexing (SCM) is a method for combining (multiplexing) many different communications signals so that they can be transmitted along a single optical fiber.

of interworking with software-defined radio (SDR) for orchestration and slicing of fronthaul resources. A DSP-assisted analogue fronthaul link is shown (for downlink only in this case, for simplicity) in Figure 27. Proper placement of signals at the DU in the multiplexing and DSP functions, can significantly simplify the de-mux and DSP functions at the RRU. Although the RRU ADC will require a high sampling rate and analogue bandwidth, methods for reducing these are proposed without negating the flexibility required of the system.

An evaluation of the DSP-assisted analogue fronthaul transport has commenced using experimental measurements and simulation [18].

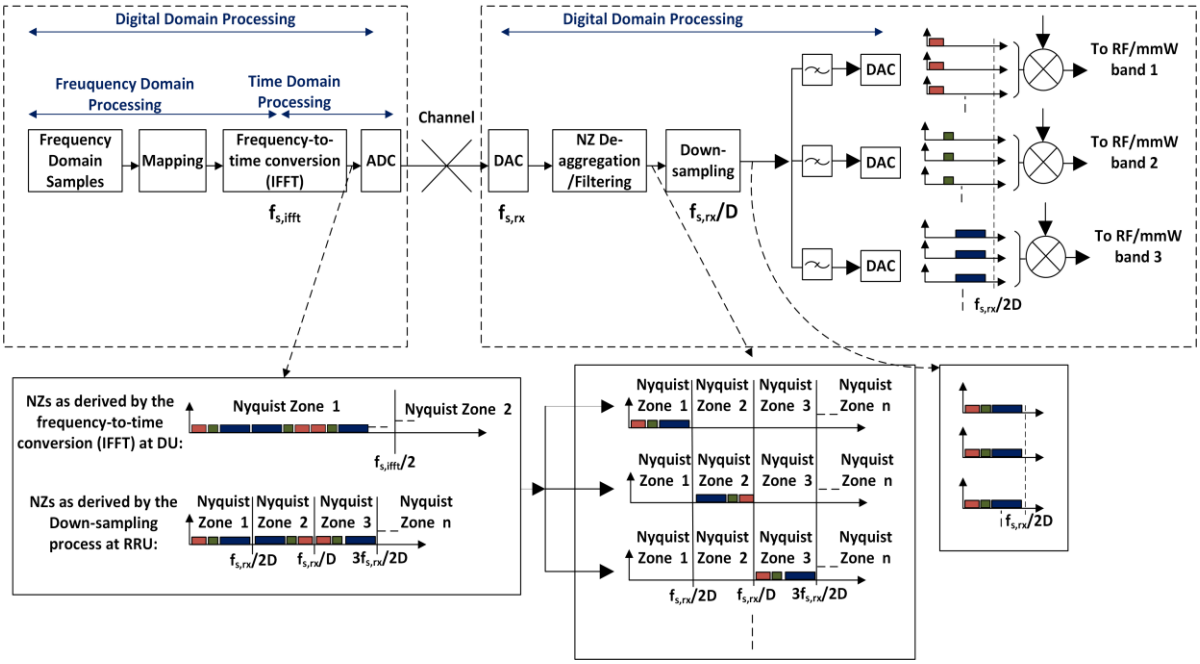


Figure 28: Conceptual view of the techniques used in the DSP-assisted RoF system.

An overview of the techniques involved in the DSP-assisted subcarrier multiplexed radio-over-fiber system is shown in Figure 28. Channels are multiplexed in the frequency domain while a single IFFT process is used to convert the multiplex into a time-domain signal with a sampling rate of $f_{s,ifft}$. An example of how this is performed is shown in Figure 29. Here, frequency domain samples, each representing a modulation-derived symbol, are packed into contiguous channels. These channels may have conjugate symmetric counterparts in the "negative" part of the spectrum, in which case they represent discrete multi-tone[51] (DMT) channels. Alternatively, both sides of the spectrum can carry independent channels in which case they represent single-sideband modulation[52] (SSB) channels. Any combination of the two channel types can be employed; for example, in Figure 29 alternating DMT and SSB channels are shown. The multiplex is distributed across two Nyquist Zones[53] (NZs), derived from the employed sampling rate ($f_{s,ifft}$), each zone with a size of $f_{s,ifft}/2$.

[51] For further information also see: *https://www.sciencedirect.com/topics/engineering/discrete-multi-tone*
[52] For more informative details also see, *inter-alia*: *https://en.wikipedia.org/wiki/Single-sideband_modulation*
[53] Also see, *among others*: *https://casper.ssl.berkeley.edu/astrobaki/index.php/Nyquist_Sampling*

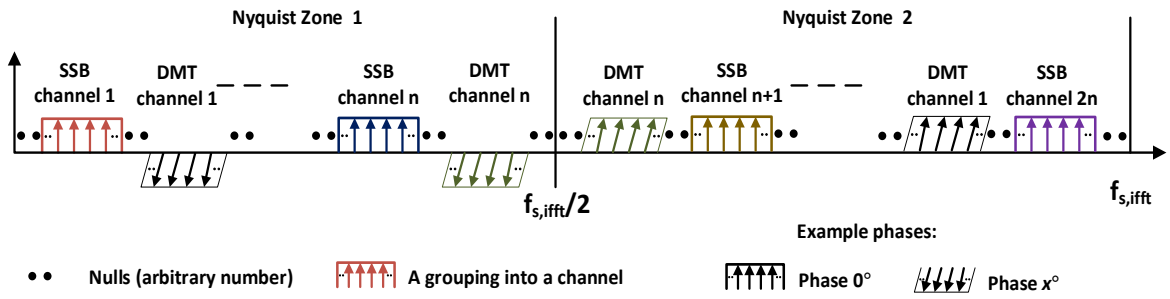


Figure 29: The process of channel creation/placement at the input of the single-IFFT stage.

Examining the remaining processes depicted in Figure 28, the placement of channels at the input of the IFFT process is not carried out arbitrarily. Instead, channels or channel groups are mapped into specific locations in the frequency domain so that they reside within NZs as derived by the sampling process at the receiver. These NZs have a size of $f_{s,rx}/D$, with $f_{s,rx}$ being the "base" sampling rate of the receiver and D the down-sampling factor used at the receiver, which is equal to the number of NZs. The mapping is carried out such that "equivalent" channels within channel groups (across the different receiver-derived NZs) are down-sampled to the same Intermediate Frequency (IF). As a result, these "equivalent" channels, which for example can be destined to a MIMO array or/and a given Radio Access Technology (RAT), can be readily up-converted to their respective mmW/RF frequency band by employing a single Local Oscillator (LO) reference signal. The processes involved in the de-multiplexing stage of the receiver are shown in more detail in Figure 30. NZ-based digital filters are used to de-aggregate channels or channel groups, followed by down-sampling by the same factor for all channels/channel groups. Then, per-channel filtering is used for each channel prior to Digital-to-Analogue Conversion (DAC) and RF/mmW up-conversion.

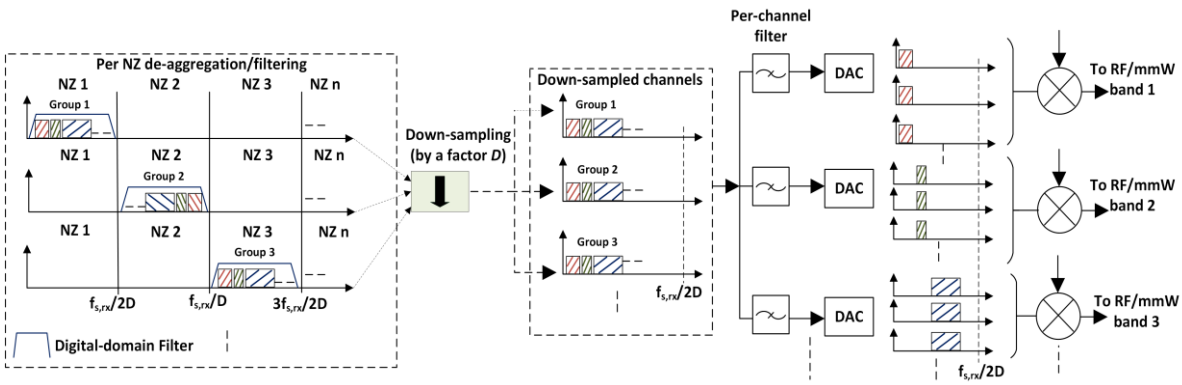


Figure 30: Conceptual view of the de-multiplexing process at the receiver.

The DSP-assisted RoF system has been experimentally verified. The experimental testbed used for performance measurements is shown in Figure 31. The optical link comprises of a Mach-Zenhder external modulator[54] (MZM), an Erbium-Doped Fiber Amplifier[55] (EDFA), a short patch of Single Mode

[54] A Mach-Zehnder modulator is used for controlling the amplitude of an optical wave. Also see, *for example*: *https://www.comsol.com/model/mach-zehnder-modulator-15061*

[55] Erbium-Doped Fiber Amplifier (EDFA) is an optical amplifier used in the C-band and L-band, where the loss of telecom optical fibers becomes lowest in the entire optical telecommunication wavelength bands. For more details see: R.J. Mears et al. (1987): Low-noise erbium-doped fiber amplifier operating at 1.54 µm. *IEEE Electronics Letters*, vol. 23, no. 19, pp. 1026–1028. Also see, *among others*: *https://www.fiberlabs.com/glossary/erbium-doped-fiber-amplifier/*

Fiber[56] (SMF) and a mmW Photodiode. The multiplex creation is carried out in Matlab[57] and the resulting time domain digitized signal is downloaded into 12 GSps Arbitrary Waveform Generator[58] (AWG). At the receive end, a 100 GHz sampling oscilloscope (OSC) is used to capture the received signal so that it can be offline processed in Matlab. The constellations, Error Vector Magnitude (EVM) and signal spectra for a 64-channel multiplex are shown in Figure 32. Note that the effective SNR is limited to a large extent by the available dynamic range of the oscilloscope and the maximum output power constraint and frequency response of the AWG[59].

The back-to-back EVM[60] ranges from 3% to 5% (the latter is for the channel at the high frequency edge of the multiplex). Still, the EVM is within the 3GPP specifications for 64 QAM. Each channel has a bandwidth of 40 MHz, resulting in a useful (i.e. excluding frequency guard bands) bandwidth of 2.5 GHz. The resulting data rate is 9.8 Gbps (16-QAM)/14.6 Gbps (64-QAM).

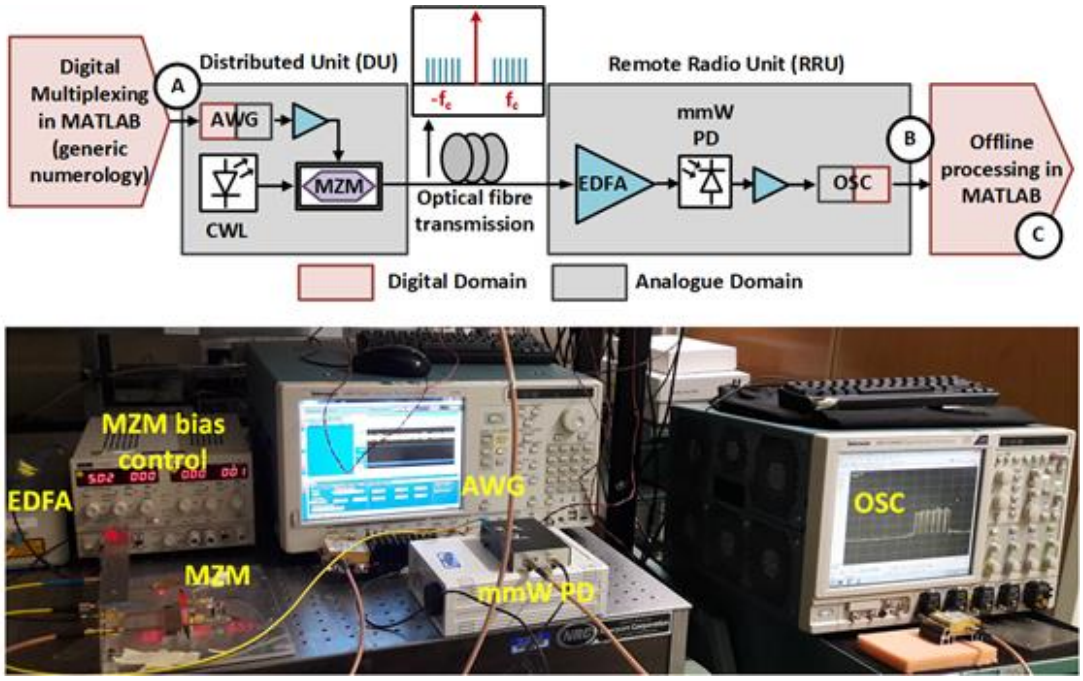


Figure 31: Experimental set-up.

[56] For more details also see, *among others*: *https://en.wikipedia.org/wiki/Single-mode_optical_fiber*

[57] For more details see: *https://www.mathworks.com/products/matlab.html*

[58] For more details also see, *among others*: *https://en.wikipedia.org/wiki/Arbitrary_waveform_generator*

[59] An arbitrary waveform generator (AWG) is a piece of electronic test equipment used to generate electrical waveforms. These waveforms can be either repetitive or single-shot (once only) in which case some kind of triggering source is required (internal or external). The resulting waveforms can be injected into a device under test and analyzed as they progress through it, confirming the proper operation of the device or pinpointing a fault in it.

[60] The error vector magnitude or EVM (sometimes also called relative constellation error or RCE) is a measure used to quantify the performance of a digital radio transmitter or receiver. A signal sent by an ideal transmitter or received by a receiver would have all constellation points precisely at the ideal locations, however various imperfections in the implementation (such as carrier leakage, low image rejection ratio, phase noise, etc.) cause the actual constellation points to deviate from the ideal locations. Informally, EVM is a measure of how far the points are from the ideal locations. Also see, *for example*: *https://en.wikipedia.org/wiki/Error_vector_magnitude*

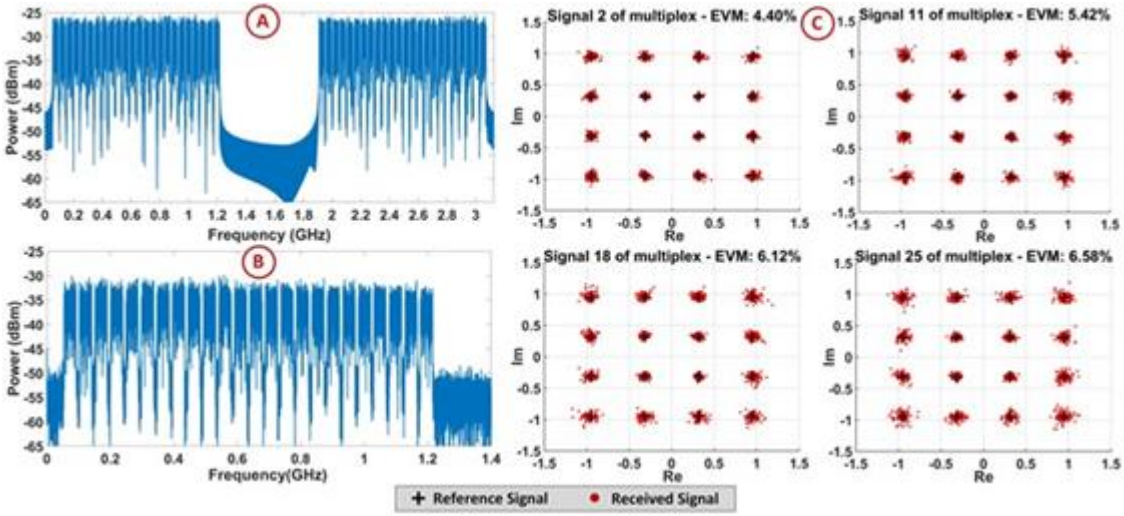


Figure 32: 64-channel multiplex: A. Transmit spectrum; B. Received one-sided spectrum; C. Received signal constellation for a channel selection.

3.3.1 Up-conversion to mmW frequencies

The experimental setup used for mmW carrier generation and up-conversion is shown in Figure 33. The optical carrier from a Continuous Wave Laser[61] (CWL) is split into two parts by a 3-dB optical coupler. One part is modulated by a phase modulator, which receives an RF tone signal at a frequency (f_M) of 25.5 GHz. This results in phase-modulated sidebands which are 25.5 GHz apart. The first order, 25.5 GHz sideband is filtered using a 25-GHz DWDM[62] filter and coupled with the other part of the optical carrier. After transmission via a short length of optical fibre, the optical carrier and 25.5 GHz sideband beat at the mmW photo-diode to produce a 25.5 GHz mmW carrier. The mmW carrier is used to up-convert a selection of the channels form the 64-channel multiplex, that following the de-multiplexing process are re-downloaded and re-generated by the AWG. The resulting spectra, constellations and EVMs following RF amplification, are shown in Figure 34.

[61] See: *https://www.spilasers.com/industrial-fiber-lasers/redpower/what-is-a-continuous-wave-cw-laser/*
[62] Also see, *for example*: *https://en.wikipedia.org/wiki/Wavelength-division_multiplexing*

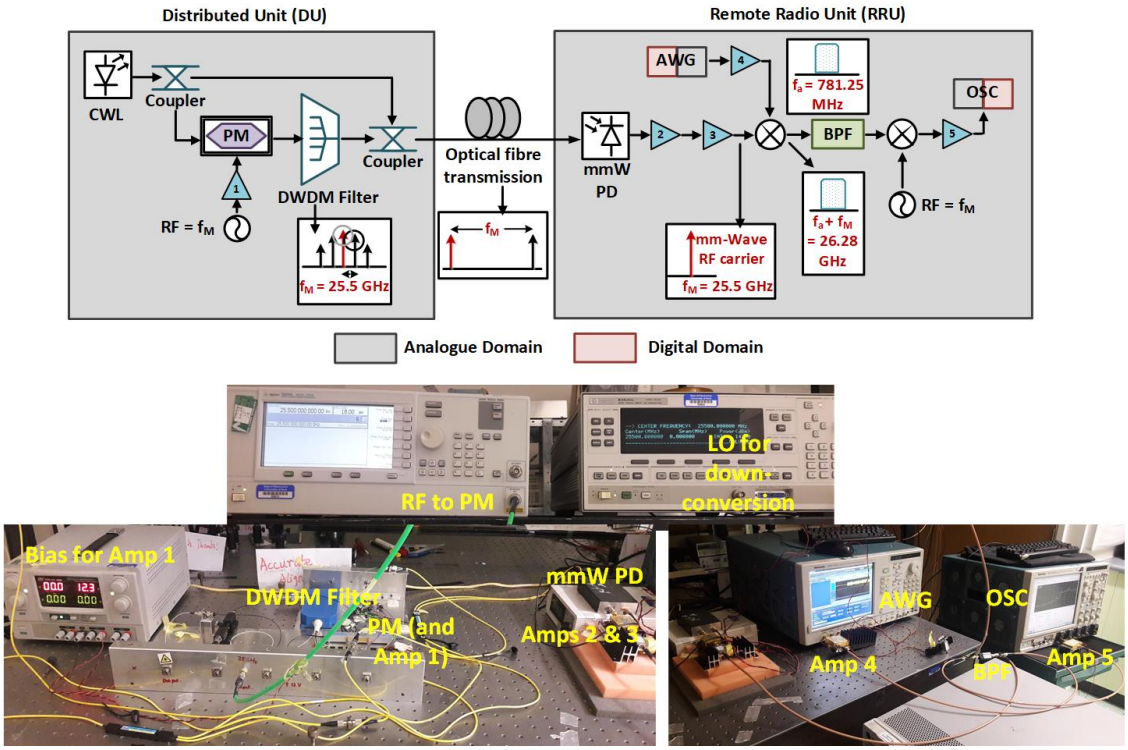


Figure 33: mmW up/down-conversion set-up.

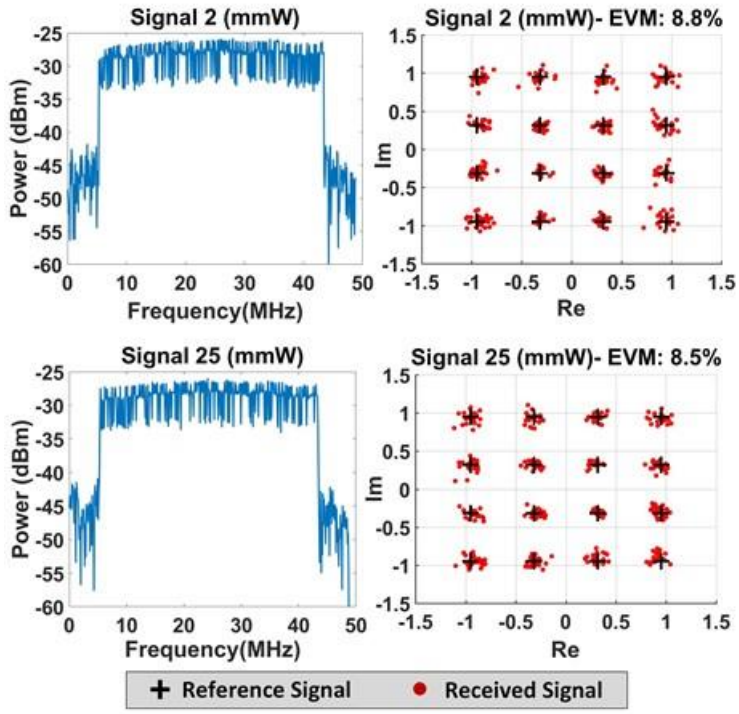


Figure 34: Received channel spectra and constellations following mmW up/down-conversion.

3.3.2 Extensions to arbitrary receiver sampling rates and analogue bandwidths

Assuming that the receiver sampling rate and analogue bandwidth capabilities specifications are constraint, there are a number of cases where the received multiplex can exceed these specifications:

- With multiplexes that have aggregate bandwidths exceeding the ADC's sampling rate specifications.
- With multiplexes that occupy sufficiently low aggregate bandwidths but are placed at an RF such that their highest frequency component exceeds the ADC's analogue bandwidth.

These limitations can be overcome by employing a Track-and-Hold Amplifier[63] (THA) at the receiver front-end with some limited analogue domain processing. This process is depicted in Figure 35. The multiplex can now be broken down into blocks of channels (sub-multiplexes) which are individually filtered (using wide-band filters) prior to being sent to the THA. The THA down-samples each block down to the same IF after which point, normal processing as described so far can resume. Note that the incorporation of the THA in effect extends the mapping technique described so far, to a mapping hierarchy. That is, whole blocks of channels (sub-multiplexes) are also mapped into NZs as derived by the THA sampling rate. The result of the THA incorporation is that the sampling rate and analogue bandwidth of the receiver can be arbitrary low. Multiplexes that exceed the sampling rate and/or analogue bandwidth capabilities of the receiver can be broken up into sub-multiplexes with a size that is within the receiver's said capabilities.

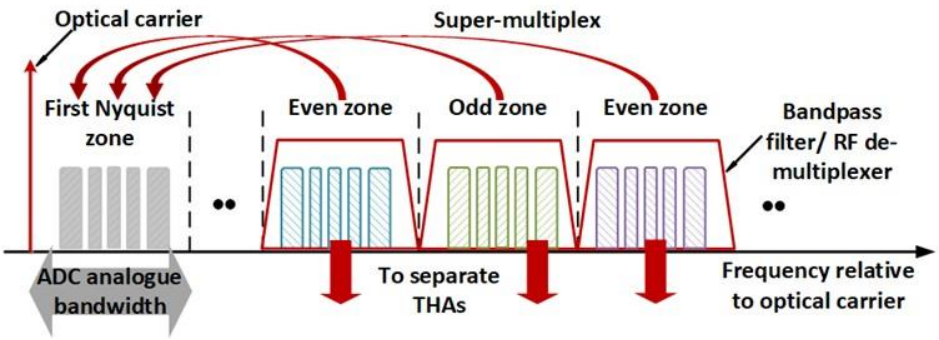


Figure 35: Incorporation of a THA with minimal analogue-domain filtering resulting in the ability to employ receivers with arbitrary low sampling rates and analogue bandwidths.

The end-to-end DSP-assisted RoF system has been modelled in Virtual Photonics Incorporated (VPI)[64] allowing for simulations with larger sampling rates than those permitted by the experimental set-up and with the incorporation of the THA. The combined THA-ADC frontend model is shown in Figure 36 with the corresponding modelling parameters. A 120-channel multiplex with mixed 5G numerologies is shown in Figure 37. Two channels, (one from each numerology) are mapped into each NZ. The channels corresponding to each numerology are generated by their own (separate) single-IFFT process while the composite multiplex is created by combining the two single-IFFT outputs in an SCM fashion, in the digital domain. One channel has a bandwidth of 20 MHz and a subcarrier spacing of 30 kHz while the other a bandwidth of 80 MHz and subcarrier spacing of 60 kHz. The receiver employs two analogue wideband filters to de-aggregate the sub-multiplexes and two 3.93 GSps ADCs, one for each sub-multiplex. Received spectra and constellations for a selection of two channels, one from each numerology, are shown Figure 38, with EVMs well within 3GPP specifications for 64QAM. The resulting data rate is 21.6 Gbps (16-QAM)/32.4 Gbps (64-QAM).

[63] See, *for example*: *https://en.wikipedia.org/wiki/Sample_and_hold.*
Also see, *among others*: *https://electronics.stackexchange.com/questions/312463/track-vs-sample-and-hold*

[64] Also see: *https://www.vpiphotonics.com/index.php*

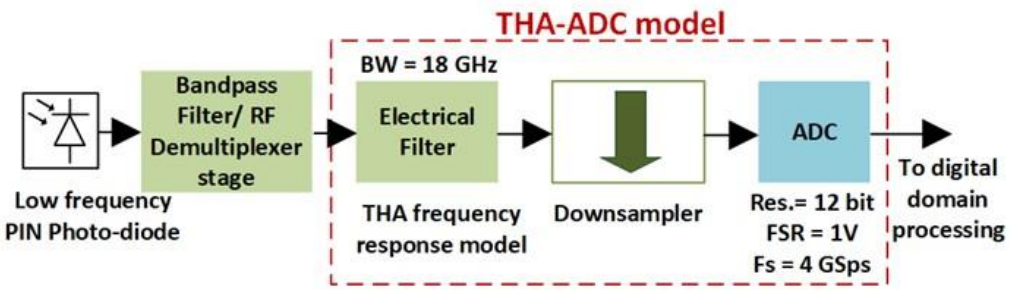


Figure 36: The model used for the receiver side front-end, now incorporating a THA.

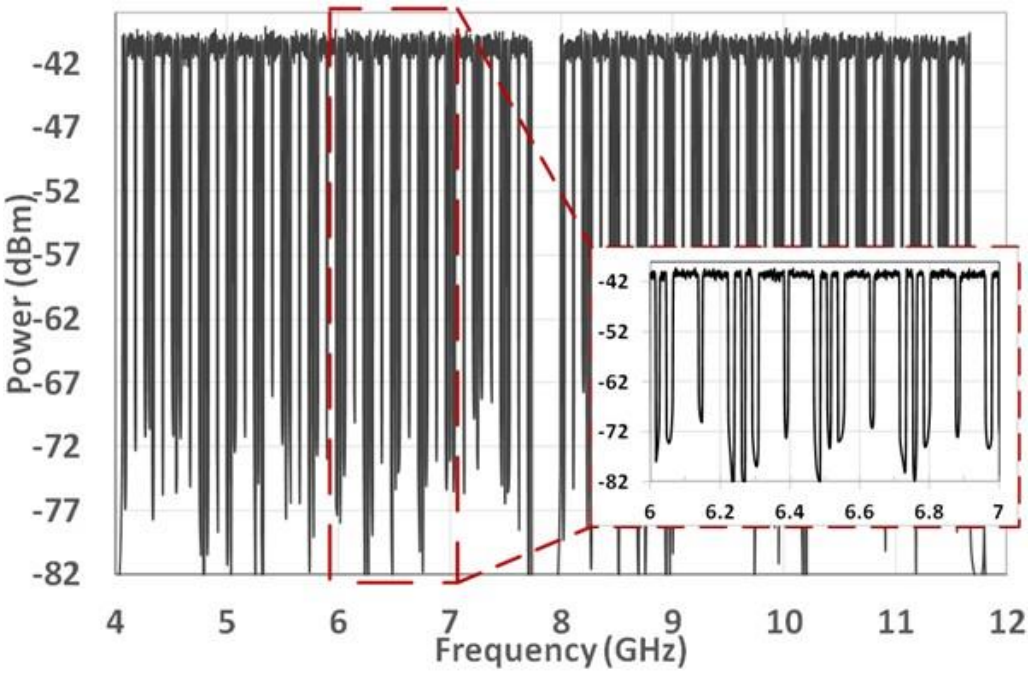


Figure 37: 120-channel multiplex with mixed 5G channel bandwidths/numerologies: Two channels per NZ with one channel having a bandwidth of 20 MHz and a subcarrier spacing of 30 kHz and the other a bandwidth of 80 MHz and subcarrier spacing of 60 kHz. The resulting data rate is 21.6 Gbps (16-QAM)/32.4 Gbps (64-QAM).

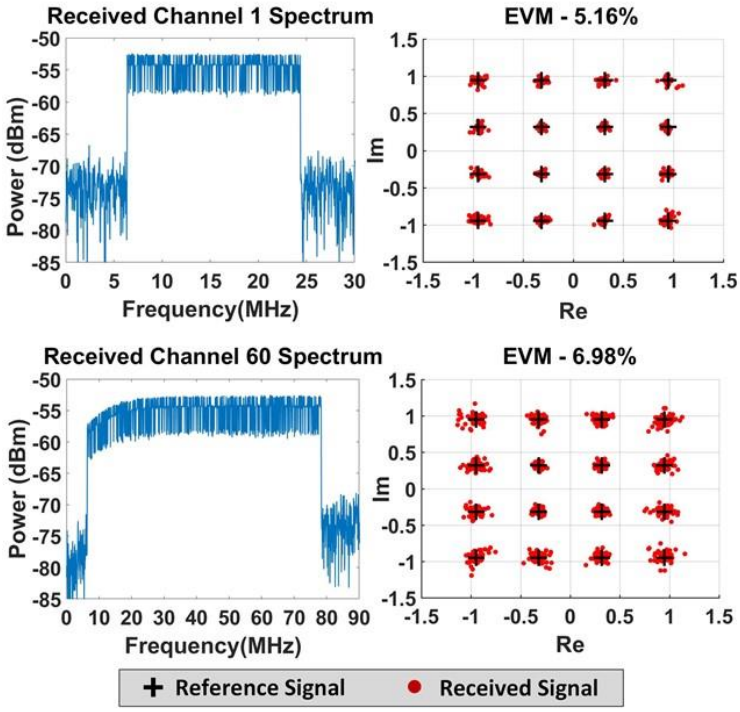


Figure 38: Received signal spectrum and constellation for a selection of channels, one form each numerology, from the 120-channel multiplex.

The multiplex size can be readily increased through scaling of the number of ADC-THA stages by the number of sub-multiplexes. A 240-channel multiplex with the same mixed 5G channel bandwidths/numerologies but different modulation schemes is shown in Figure 39. The corresponding received spectra and constellations, following 60 GHz up/down-conversion are shown in Figure 40. The resulting data rate now is 49.7 Gbps.

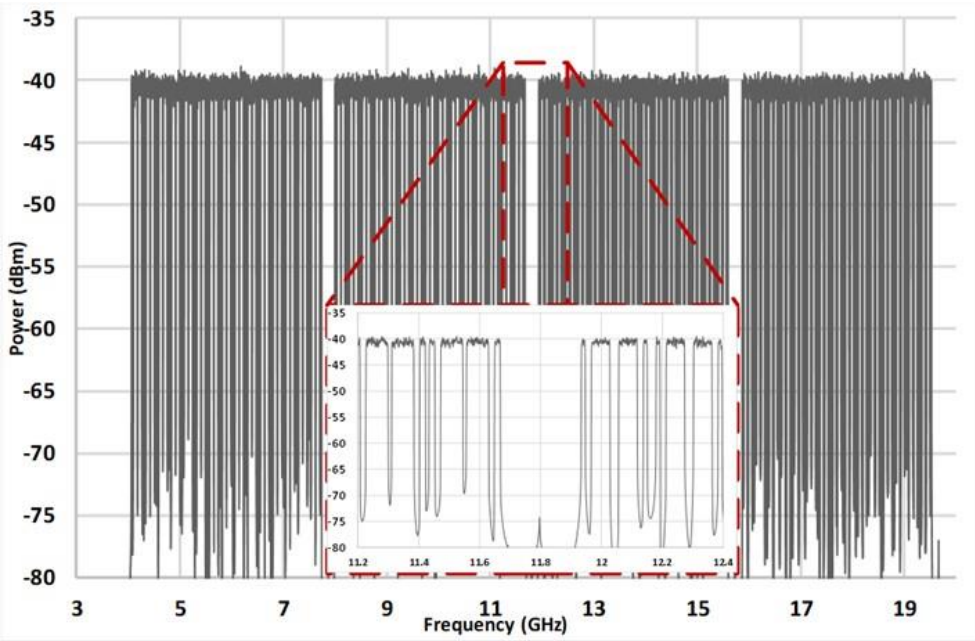


Figure 39: A 240-ch multiplex with mixed 5G channel bandwidths/numerologies and different modulation schemes: Two channels per NZ with one channel having a bandwidth of 20 MHz and a subcarrier spacing of 30 kHz and the other a bandwidth of 80 MHz and subcarrier spacing of 60 kHz. The resulting data rate is 49.7 Gbps.

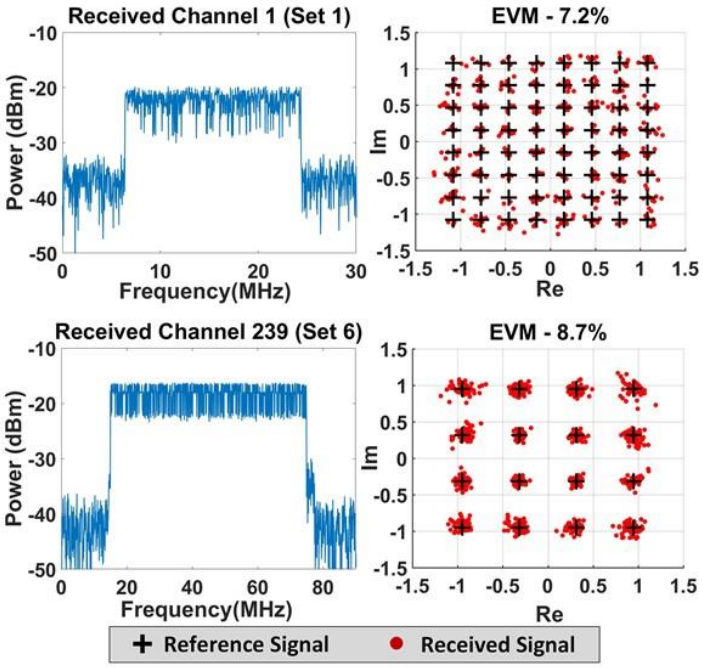


Figure 40: Received signal spectrum and constellation for a selection of channels, one form each numerology, from the 240-channel multiplex following up/down-conversion to/from 60 GHz.

3.4 Function virtualisation and SDN

In order for the new fronthaul to operate within an end-to-end network slicing and orchestration framework, the SDN control of fronthaul functions and operation is being examined [19]. The Gb Ethernet switches are made SDN-controllable; in Figure 41, one switch/aggregator is shown as an SDN switch, reflecting the current testbed, although in the future, all switches/aggregators in the fronthaul may be. A probing server is used to gather the results from in-line hardware probes at measurement ports across the fronthaul (and xhaul, generally), passing analysed results to a KPI aggregation and SDN controller server. This server may also send modifications of what it requires monitoring to the probing server.

OpenFlow messages are sent to/from the SDN controller and SDN switches to modify, *for example*, queuing disciplines for different flows/slices. A GPS time server also uses the probing server to distribute timing information (Precision Time Protocol (PTP) packets) to the fronthaul network nodes.

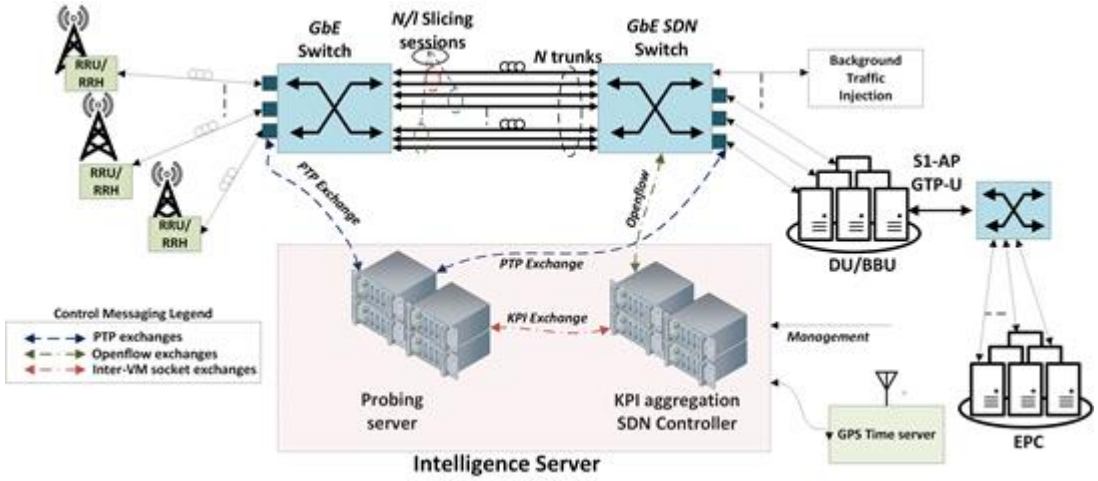


Figure 41: SDN in an Ethernet fronthaul.

In 42, a closer look is taken at the reconfiguration possibilities for the fronthaul through using SDN. Trunk links are assumed between Ethernet switch/aggregation nodes, with multiple physical ports and VLANs being used to separate different flows. The processing server and SDN controller virtual machines (VMs) may take KPIs from over-the-air measurements and from measurements on the fronthaul using the in-line hardware probes. Based on these and the assignment of flows to different VLANs and Classes of Service-CoSs (based on the flow requirements), the SDN controller is able to instruct switches/aggregators in such a way as to carry out load balancing between different physical ports/links, traffic steering (e.g. moving certain traffic away from a trunk to protect the Quality of Service (QoS) requirements of other flows on that trunk) and flow/slice isolation. This work is being extended to develop programming interfaces that can be used by network management and orchestration functions.

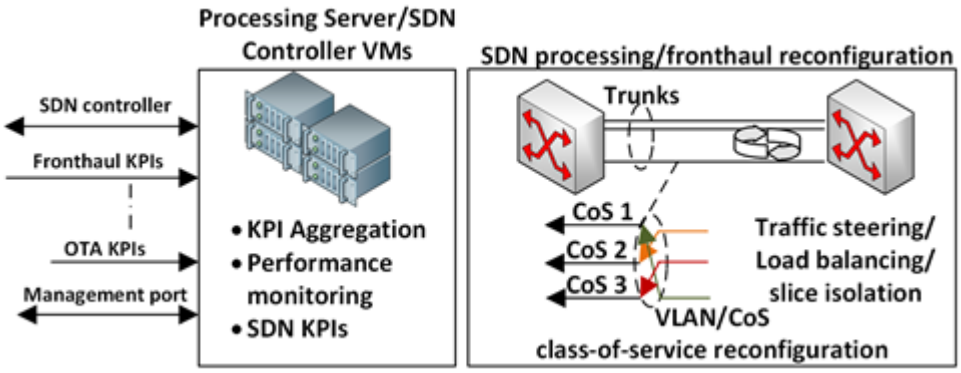


Figure 42: SDN reconfiguration possibilities for an Ethernet fronthaul.

An example of Ethernet fronthaul monitoring and SDN processing is shown in Figure 43. Here a number of flows (specifically DLSCH and DCI flows) originating from the Option-6 (MAC/PHY) split are differentiated based on VLAN tags. These flows contend with background traffic. One-way latency measurements are used to inform an SDN controller, which once a threshold is exceeded, applies traffic steering, steering (or path switching) the background traffic. As a result latency (and latency variation) of all flows is reduced significantly.

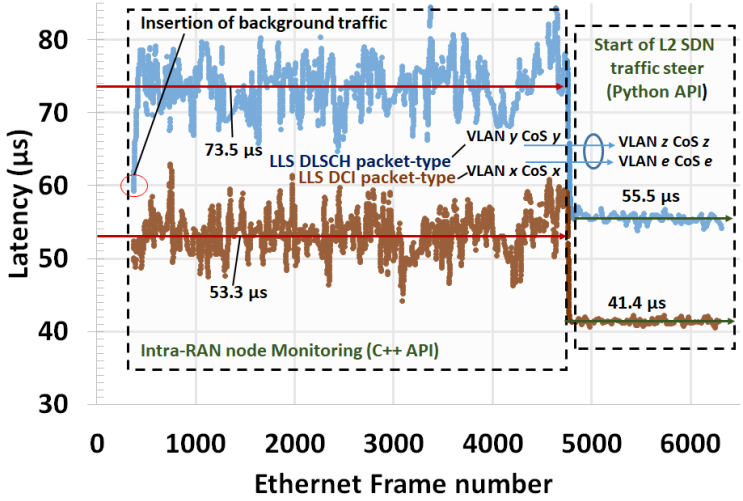


Figure 43: SDN traffic steering within an evolved fronthaul (MAC/PHY LTE functional split).

4 Network virtualization and slicing interoperability

4.1 Network slicing concept

4.1.1 General introductory framework

Network slicing enables the creation of parallel virtual telecommunication networks (which in some cases can be integrated with applications) over a common distributed cloud infrastructure [20]. The main advantage of this approach is the ability of on-demand creation of isolated networking solutions, which are combined or tailored for specific applications and give slice management capabilities to slice tenants. Most of the approaches to network slicing follow the concept described by Next Generation Mobile Networks[65] (NGMN). The "key enabler" of network slicing is the European Telecommunication Standardisation Institute[66] (ETSI) Network Functions Virtualisation (NFV) framework [21]. A network slice is an end-to-end logical network that can be created on demand and customized according to the needs of specific applications. The creation of a slice can be triggered not only by a telco operator or 3rd parties (tenants) but in some cases also by the end-users. The concept is seen as one of the key foundations of the 5G network; however, it should be considered as a universal and basic technique for building future networks and services, which has mainly three main benefits:

- low Capital Expenditures (CAPEX) of deployment of such networking solutions combined with the efficient use of common infrastructure resources;
- ability to create a slice customized for a specific service, and;
- slice ownership and its management rights – the slice operator can see the slice as its own network and can manage it.

The network slicing concept is currently a topic of numerous research projects, mostly linked with 5G. The projects have provided some progress, but there is no integrated network slicing approach that combines the research efforts into a single, coherent concept. Moreover, many details of the proposed approaches still have to be developed. So far, there is no standardized approach to network slicing that addresses the typical carrier-grade requirements as interoperability, scalability, controllable performance, security, accounting and more.

4.1.2 Network Slicing in the 5G digital era

Digital transformation, strongly assisted by the power of mobility, cloud and broadband, is actually realised within all modern industries. This sort of transformation supports the establishment of a great variety of innovative uses cases (UCs) for the end-users/consumers[67] (either corporate or residential) upon the basis of the already deployed 4G/LTE network infrastructures. The challenges and the opportunities become much greater if consider the actual trends for realising an effective development towards the 5G era. In order to deal with these opportunities at scale and at targeted cost/investments, requests for efficiency and flexibility implicate for the definition, establishment and validation of specific network capabilities, able to support expected views. In this context, the

[65] NGMN main website: *https://www.ngmn.org/home.html* [Accessed April 2019].

[66] ETSI – Welcome to the World of Standards*! https://www.etsi.org/* [Accessed April 2019].

[67] The possibility of tailoring mobile network properties to the needs of the business through the configuration of a large set of parameters offers unsurpassed flexibility. However, with such a diverse range of possible requirements from verticals, operators will need to manage risks from excessive complexity in the service offering and cumbersome management, leading to higher costs.

network slicing[68] is an innovative technology that fulfills this challenge, permitting multiple logical networks to be created on the top of a common shared physical infrastructure.

The key benefits include[69]: (i) Greater elasticity, robustness, secure and stable operations and functionalities through the compartmentalisation of the network, applied end-to-end (E2E); (ii) uncompromising and customizable slices, each optimized for the needs of the services- or segment cluster they are defined to serve, and; (iii) built-in flexibility and efficiency with Artificial intelligence (AI)-powered automated service orchestration, from test to launch to the maintenance of new services.

In the scope of the efforts towards achieving the much promising 5G potential, network slicing responds, explicitly, to any request about how to perform both increased efficiencies and revenues, through differentiation and faster time-to-market.

Network slicing, in its simplest description, is the capability to modify a set of functions to improve use of the network for each involved mobile device. All of the functionality needed is accumulated so that to optimize that device's ability to find the correct network, access the network efficiently and securely, and be attached to the core network (CN) with the set of functionalities needed by that device. From a business point of view, a slice includes a combination of all the relevant network resources, functions, and assets required to fulfil a specific business case or service, including OSS (Operations Support Systems), BSS (Business Support Systems) and DevOps processes.

Network slicing[70],[71] refers to partitioning of one physical network into multiple virtual networks, each architected and optimized for a specific application/service. Specifically speaking, a network slice is a virtual network that is created on top of a physical network in such a way that gives the illusion to the slice tenant of operating its own dedicated physical network. A network slice is a self-contained network with its own virtual resources, topology, traffic flow and provisioning rules. The related customizable network capabilities include data speed, quality, latency, reliability, security, and services. These capabilities are always provided based on a Service Level Agreement (SLA) between the mobile operator and the business customer.

There may be various network slices to meet the specific communication needs of different users in the future mobile network systems. For example, a massive industrial IoT slice may need a light 5G core, no handover but a large number of connections. On the other hand, a mobile broadband slice may need a high capacity core, full feature mobility and low latency. Slices are logically isolated, but resources can be shared among them. A network slice comprises of dedicated and/or shared resources, e.g. in terms of processing power, storage, and bandwidth and has isolation from the other network slices.

Network slicing[72] allows core networks to be logically separated, with each slice providing customized connectivity, and all slices running on the same, shared infrastructure or on separate infrastructures as the operator desires, following to related market needs. This is a much more flexible solution than a single physical network providing a maximum level of connectivity. Based on recent technological

---

[68] For more details also see, *among others*, the detailed discussion provided in: *https://www.ericsson.com/en/digital-services/trending/network-slicing?gclid=EAIaIQobChMIobWOmKWn4wIVD853Ch3xTA5cEAAYASAAEgLfovD_BwE*

[69] J. Elliott, and S. Sharma (2016, July): "Dynamic End-to-End Network Slicing unlocks 5G Possibilities". Nokia. Available at: *https://www.nokia.com/blog/dynamic-end-end-network-slicing-unlocks-5g-possibilities/*

[70] Also see: Global System for Mobile Communications Association (GSMA) (2017, November): "An Introduction to 5G Network Slicing". Available at: *https://www.gsma.com/futurenetworks/wp-content/uploads/2017/11/GSMA-An-Introduction-to-Network-Slicing.pdf*

[71] S. Sharma, R. Miller, and A. Francini (2017, August): A Cloud-Native Approach to 5G Network Slicing. *IEEE Communications Magazine*, vol. 55, no. 8, pp. 120-127.

[72] Also see, among others: H.-J. Einsiedler, A. Gavras, P. Sellstedt, R. Aguiar, R. Trivisonno, and D. Lavaux (2015): "System design for 5G converged networks". In IEEE (Ed.), Proceedings of the 2015 European Conference on Networks and Communications (EuCNC-2015), pp. 391-396. Paris, France, June 29 - July 02, 2015.

trends coming from the 5G growth[73], it appears that networks will need to be deployed using different hardware technologies, with different feature sets placed at different physical locations in the network, depending on the use case[74]. To support a specific set of services efficiently, a network slice should have access to different types of resources, such as infrastructure -including VPNs, cloud services[75] and access- as well as resources for the core network in the form of VNFs. The flexibility of 5G core networks will improve significantly by supporting a full separation of control plane and user plane, and through adopting selected SDN principles and technologies[76].

Network slicing will allow 5G networks to be sliced logically into multiple virtual networks[77]. Each slice can be optimized to serve a specific vertical application to efficiently support network services, thus providing high degree of flexibility in enabling several use cases to be active concurrently. This is already a well understood methodology in the wireless industry in some limited environments, such as software-defined core networks[78]. It provides a Network as-a-Service (NaaS) model, which can be very flexible to allocate and reallocate resources according to the dynamic demands such that it can customize network slices for diverse and complex 5G communication scenarios.

Network slicing will be the fundamental feature of 5G networks. Slice-based 5G has the following significant advantages when compared with the traditional networks: (i) Network slicing can provide logical networks with better performance than one-size-fits-all networks; (ii a network slice can scale up or down as service requirements and the number of users change; (iii) Network slices can isolate the network resources of one service from the others. The configurations among various slices do not affect each other; therefore, the reliability and security of each slice can be enhanced; (iv) Finally, a network slice is customized according to service requirements, which can optimize the allocation and use of physical network resources.

Network slicing influences the up-to-date advances in cloud mobile access and core. By combining cloud technologies with the capabilities of software defined networking (SDN) and network function virtualisation (NFV)[79], this delivers the essential tools to realize network slicing[80]. Virtualisation technologies offer a fundamental basis for network slicing by allowing use of both physical and virtual resources to generate the service they are intended for. Virtualisation can enable separation

---

[73] J.G. Andrews, S. Buzzi, W. Choi, S.V. Hanly, A. Lozano, A.C.K. Soong, and J.C. Zhang (2014, June): What will 5g be? *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1065-1082.

[74] For further related information also see, *among others,* the discussion and the context proposed in: K. Makhijani, J. Qin, L. Geng, L. Qiang, S. Peng, X. de Foy, A. Rahman, A. Galis, and G. Fioccola (2017, October): "Network Slicing Use Cases: Network Customization and Differentiated Services", Internet Engineering Task Force (IETF). Available at: *https://datatracker.ietf.org/doc/draft-netslices-usecases/*

[75] Operators are already in the midst of virtualizing their networks and services, in the form of cloud services and software-defined networking (SDN)/network functions virtualisation (NFV) - paving the way to further evolution of the backhaul. Essentially, the 5G architecture will incorporate an all-cloud digital transformation of networks, operation systems and related services.

[76] M. Reddy Sama, X. An, Q. Wie, and S. Beker (2016): "Reshaping the Reshaping the mobile core network via function decomposition and network slicing for the 5G Era". In IEEE (Ed.), Proceedings of the 2016 IEEE Wireless Communications and Networking Conference. Doha, Qatar, April 03-06, 2016.

[77] A device that offers enhanced functionality, such as a smartphone, can find the radio interfaces that it need. The device can be authenticated and attached to a diverse set of network slices that are each tailored to a specific purpose such as: streaming video, voice calls, internet browsing, chatting and so on. Such devices may need to carry on high-quality voice and video calls while, simultaneously continuing background browsing and upload/download functions.

[78] For further information also see, among others: I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck (2018): Network slicing and softwarization: A survey on principles, enabling technologies, and solutions. *IEEE Communications Surveys and Tutorials,* vol. 20, no. 3, pp. 2429-2453.

[79] Also see, *among others*: Open Networking Foundation (ONF) (2015, October): TR-518: "Relationship of SDV and NFV - Issue 1". Available at: *https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/onf2015.310_Architectural_comparison.08-2.pdf*

[80] Also see, *inter-alia*: J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J.J. Ramos-Munoz, J. Lorca, and J. Folgueira (2017, May): Network Slicing for 5G with SDN/NFV: Concepts, Architectures and Challenges. *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80-87.

of the software from the hardware and offer the possibility to instantiate many functions on a common infrastructure. With this approach, the infrastructure can be shared by different tenants and provide different services. It is proposed that this trend will continue, where virtualisation technologies will be applied across RAN and portable services and across portable devices to wearable devices.

5G core network slicing can provide tailored logical networks for different services or verticals in an agile and adaptable fashion. The key technologies of core network slicing are SDN and NFV. SDN technology can be used to separate the control plane and data plane of the core network so that the control plane and data plane can be deployed independently. The control plane can be centralised in order to facilitate management, while the data plane can be distributed. For example, the data plane in a network slice for low-latency services can be distributed on the network edge in combination with the Mobile edge Computing (MEC) technology[81],[82]. The NFV technology can provide the necessary virtual network functions for the data plane of the core network slices according to the service type. These virtual network functions can be scaled on-demand as the service changes dynamically.

When examining the way how network slicing is supported in the RAN[83], it is important to take into account two essential consider two features, that is: (i) The radio access type (RAT) that supports the network services provided by the slice, and; (ii) the configuration of RAN resources to appropriately interface with and support the network slice. From the RAN point of view, the main task of RAN slice is to achieve the sharing and flexible management of spectrum resources. Therefore, the software-defined RAN is a promising approach to implement RAN slicing. The sharing of radio resources between different slices can be accomplished by scheduling conducted by the controller in software-defined RAN[84]. The controller can allocate radio resources to a slice according to service request characteristics such as traffic load. Moreover, different radio access methods for RAN slices are needed in terms of diverse service requirements. For example, besides the mobile broadband access in the traditional cellular networks, the critical communication slice in 5G needs low-latency radio access. In addition, the implementation of RAN slicing needs logical decompositions of RAN functions to figure out which functions are common functions and which functions are provided with specific hardware[85].

The 5G has to fulfil several multi-dimensional KPI goals[86]. Some of the goals cannot be met simultaneously; for example, low latency and reliability often come at the cost of spectral efficiency. These goals implicate for, among others: (i) Optimization for low latency and high reliability; (ii)

---

[81] Advanced 5G services are expected to be provided at the network edge so as to reside "much closer to the user" in order to enhance delay and perceived performance, for example, by adopting the ETSI MEC paradigm. Therefore, a flexible service chaining should also be improved to incorporate dynamic services considering edge network locations and might be combined with corresponding VNFs to ensure a joint optimization of services and networking operations. Edge server locations can also be exploited for storage, computation, and dynamic service creation within a given network slice by verticals and over-the-top providers, introducing another multi-tenancy dimension. For further reading about MEC also see, *among others*: *http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing*

[82] S. Kekki, W. Featherstone, Y. Fang, P. Kuure, A. Li, A. Ranjan et *al.* (2018, June): ETSI White Paper No. 28: "MEC in 5G Networks - First Edition". European Telecommunications Standards Institute (ETSI). Available at: *https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf*

[83] See, *among others*: A. Ksentini and N. Nikaein (2017, June): Toward Enforcing Network Slicing on RAN: Flexibility and Resources Abstraction. *IEEE Communications Magazine*, vol. 55, no. 6, pp. 102-108.

[84] H. Xiang, W. Zhou, M. Daneshmand, M. Peng (2017, December): Network Slicing in Fog Radio Access Networks: Issues and Challenges. *IEEE Communications Magazine*, vol. 55, no. 12, pp. 110-116.

[85] E. Westerberg (2016): "4G/5G RAN architecture: how a split can make the difference", *Ericsson Technology Review*, vol. 93, no. 6, pp. 1-16. Available at: *https://www.ericsson.com/media-cdn/49ec87/siteassets/publications/ericsson-technology-review/docs/2016/etr-ran-architecture.pdf*

[86] International Telecommunication Union - Radiocommunications Sector (ITU-R): Recommendation ITU-R M.2083: IMT Vision - "Framework and overall objectives of the future development of IMT for 2020 and beyond", September 2015. Available at: *https://www.itu.int/rec/R-REC-M.2083*

optimization for massive connection; (iii) optimization for high spectrum efficiency; (iv) optimization for high data rate, and; (v) optimization for wide coverage. New features have been included in 3GPP LTE later releases for supporting services such as narrow band internet of things (NB-IoT) and machine type communications (MTC). In addition, configuration parameters have been added to be able to create direct device-to-device (D2D) communications. The LTE radio architecture was primarily designed for MBB services, so co-existence with legacy LTE radio is required when adding new RATs, which imposes constraints on the new RAT design. In fact, RAN slicing[87] is a mapping of slice-ID to a set of configuration rules for the RAN[88]. So rather than separate control and user plane RAN functions on a per-slice basis, there are RAN configuration rules associated with each slice to accomplish the network services supported by the slice.

Some design and operational requirements on RAN slicing are: (i) Each slice is supported in the RAN by applying a set of configuration rules to the RAN control and user plane functions; (ii) some network functions are common to several slices (e.g. mobility management), and; (iii) common control functions coordinate RAN resource usage among the slices. This is to improve the performance and efficiency of the entire 5G system.

The following design aspects can be considered in RAN slicing: (i) resource management; (ii) control plane (CP) and user plane (UP); (iii) slice-specific admission control, and; (iv) user equipment (UE) awareness on the RAN configurations.

Network slicing implementation is end-to-end from the core through the RAN. In the core network[89], NFV and SDN virtualise the network elements and functions in each slice to meet its own requirement. In the RAN, slicing can be built on physical radio resources (e.g. transmission point, spectrum, time) or on logical resources abstracted from physical radio resources[90].

4.1.3 Actual trends for further growth

International fora such as NGMN[91] and 3GPP[92] have developed the definition and use cases[93] for network slicing so that the SDOs (Standards Development Organizations) can provide detailed studies to understand the features and functionalities that will be required for network slicing beyond what is already defined in 3GPP Rel-13[94] and ETSI NFV[95]. More specifically, the 3rd Generation Partnership

---

87 X. Foukas, G. Patounas, A. Elmokashfi, and A.K. Marina (2017): Network slicing in 5G: Survey and challenges. *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94-100.

88 Also see, for example: I. da Silva, G. Mildh, A. Kaloxylos, P. Spapis, E. Buracchini, A. Trogolo, G. Zimmermann, and N. Bayer (2016): "Impact of network slicing on 5G radio access networks". In IEEE (Ed.), Proceedings of the 2016 European Conference on Networks and Communications (EuCNC-2016), pp 153-157. Athens, Greece, June 27-30, 2016.

89 See, *for example*: P. Rost, A. Banchs, I. Berberana, M. Breitbach et *al.* (2016, May): Mobile network architecture evolution toward 5G. *IEEE Communications Magazine*, vol. 54, no. 5. pp. 84-91.

90 For further informative details also see, *among others*: Y.-i. Choi, and N. Park (2017): "Slice Architecture for 5G Core Network". In IEEE (Ed.), Proceedings of the 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN-2017). Milan, Italy, July 04-07, 2017.

91 Next Generation Mobile Networks (NGMN) Alliance (2016, February): "*Description of Network Slicing Concept by NGMN Alliance- Version 1.0"*. Available at: *https://www.ngmn.org/fileadmin/user_upload/160113_Network_Slicing_v1_0.pdf*

92 The Third Generation Partnership Project (3GPP): 3GPP TS 22.891 v14.0.0, "Feasibility Study on New Services and Markets Technology Enablers; Stage 1", Release 14. Available at *https://portal.3gpp. org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2897*

93 For example, the Next Generation Mobile Networks (NGMN) Alliance's 5G White Paper alone (accessible at: *https://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf*) alone envisions 28 use-cases combined with multi-Radio Access Technologies (RATs).

94 Release 13 comprises around 170 high-level features and studies. In addition to enhancements to existing services and features, this release saw the completion of the first set of specifications covering mission-critical services, in particular mission-critical Push-To-Talk (MCPTT), the essential functionality for LTE to be used by "blue light" services for private mobile radio voice communication. A study was also completed on "isolated E-UTRAN operation" (i.e. mobile-to-

Project (3GPP) has identified network slicing as one of the key technologies to achieve a multiplicity of goals in future 5G networks. Some 3GPP work items have the features of network slicing; *for example*, the dedicated core (DÉCOR) feature supports the operator to deploy multiple dedicated core networks by sharing a common "Public Land Mobile Network (PLMN)"[96].

The requirements for 5G network slicing have initially been presented in clause 5.2.3 of 3GPP Technical Report 22.891. These have provided a robust, high-level outlook and direction for the next generation (NG) work of 3GPP with respect to network slicing:

- The operator shall be able to create and manage network slices that fulfil required criteria for different market scenarios.
- The operator shall be able to operate different network slices in parallel with isolation that for example, prevents one slice's data communication to negatively impact services in other slices.
- The 3GPP system shall have the capability to conform to service-specific security assurance requirements in a single network slice, rather than the whole network.
- The 3GPP system shall have the capability to provide a level of isolation between network slices to confines a cyber-attack to a single network slice.
- The operator shall be able to authorize third parties to create, manage a network slice configuration (e.g. scale slices) via suitable APIs, within the limits set by the network operator.
- The 3GPP system shall support network slice elasticity in terms of capacity with no impact on the services of this slice or other slice.
- The 3GPP system shall be able to change the slices with minimal impact on the ongoing subscriber's services served by other slices: specifically, of new network slice addition, removal of existing network slice or update of network slice functions or configuration.
- The 3GPP System shall be able to support end-to-end resource management for a network slice.

Slicing technology is evolving rapidly. In the process, it is drastically changing the architecture and nature of communications. Many use cases[97] are emerging that have diverse requirements in terms of speed, number of connections, availability, reliability, battery life and latency. These use cases may be broadly categorized as Mobile Broadband (MBB), Machine Type Communication (MTC) and critical MTC.

Applications like remote operation of machinery, tele-surgery and smart metering all necessitate for connectivity but with enormously different characteristics. New technologies such as virtualisation, network programmability and network slicing enable logical networks that are customized to "meet" the needs of each separate application. Network slicing provides greater insight into network

---

mobile) for public safety use, and the architecture for supporting emergency services over wireless local area networks was investigated. Work continued on security issues to ensure that new services are free from the threat of hacking, denial of service attacks etc. For the first time, serious investigations were conducted into network virtualisation and how this might be realised in the context of a 3GPP network. For more details see: *https://www.3gpp. org/release-13*

95 European Telecommunications Standards Institute (ETSI): ETSI GS NFV 002 v1.2.1 (2014-12): "Network Functions Virtualisation (NFV); Architectural Framework". Available at: *https://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.02.01_60/gs_nfv002v010201p.pdf*

96 An exhaustive description about how the next generation wireless system will support network slicing is given in: 3GPP TR 23.799 V14.0.0 (2016, December): "Study on Architecture for Next Generation System (Release 14)", found at: *https://portal.3gpp. org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3008*

97 Global System for Mobile Communications Association (GSMA) (2018, April): "Network Slicing: Use Case Requirements". Available at: *https://www.gsma.com/futurenetworks/wp-content/uploads/2018/04/NS-Final.pdf*

resource utilisation, with each customized network slice, to match the level of delivery complexity required by the supported services. As a result, new products and services can be brought to market rapidly and be easily adapted to fast-changing demands. Through appropriate solutions, network slices can be set up based on various service characteristics, such as bandwidth- and latency demands. One use case category example with specific characteristics is Augmented Reality (AR), requiring high throughput, data speed, and low latency. Another use case is Massive IoT such as connectivity for smart electricity meters, requiring high-reliability data - only at given data rates and security levels. A third use case category is capacity or coverage on demand, such as is the case in an emergency, for mission-critical services.

Designing a network that can simultaneously support both a wide variety of use cases and demanding performance requirements, all with a single set of standard network functions, would be extremely complex and prohibitively expensive. The alternative can be the option "network slicing", which is considered to be "key" for meeting 5G's diverse requirements, including future-proof scalability and flexibility.

The network slicing concept enables the network elements and functions to be easily configured and reused in each network slice to meet a specific requirement[98]. The implementation of network slicing is conceived to be an end-to-end feature that includes the core network and the RAN. Each slice can have its own network architecture, engineering mechanism and network provisioning

5G RAN slicing can be implemented through logical abstraction of physical radio resources such as spectrum and physical hardware such as a base station. SDN and NFV can configure the virtual network resources flexibly, which include network bandwidth, server processing capability, network element processing capability, etc., to build the core network slices for specific service requirements. RAN slices and core network slices can either be dedicated to a class of service users or shared between multiple classes of service users. In order to provide an identical slice for a specific service, a common function "slice selecting function" is needed. It is responsible for selecting RAN and core network slices to form the end-to-end slice.

Operators will need to invest in new technologies to "address" the efficiency and flexibility demands of these new service deployments, if they are to make the most of the potential value that can be generated. Network slicing is one solution that has been emerged and provides the capability to enable new business models across a wide range of industries[99]. It allows operators to segment the network to support particular services and deploy multiple logical networks for different service types over one common infrastructure. A "service type" can be defined as a unique set of services requiring different network design and validation efforts. On the other hand, service operators can utilize or benefit from Network Slicing through multi-tenancy, enabling different customized network infrastructures for different group of services across different network domains and operating them independently. 5G networks must support a variety of very diverse use cases with different requirements for latency, throughput, and availability. The dynamic network slicing concept offers a way to optimize 5G networks to address all use cases efficiently.

Dynamic network slicing allows for the design, deployment, customization, and extended optimization of different network slices running on a common network infrastructure. It leverages a great variety of innovations in (cloud) mobile access and core. Essentially, it is realised via the exploitation and the proper use of SDN, NFV, end-to-end orchestration, network applications and services[100], and analytics[101]. Thus, dynamic network slicing can become one of the 5G era tools and

---

[98] See, *for example*: X. Li, M. Samaka, S. Anthony Chan, D. Bhamare, L. Gupta, and R. Jain et *al.* (2017, September): Network Slicing for 5G: Challenges and Opportunities. *IEEE Internet Computing*, vol. 21, no. 5, pp. 20-27.

[99] See, *for example*: N. Davies, and P. Thompson (2017, January): Challenges of Network Slicing. *IEEE Softwarization Magazine.* Available at: *https://sdn.ieee.org/newsletter/january-2017/challenges-of-network-slicing*

[100] Network services specify a set of resource requirements to offer desired Quality of Experience (QoE) to it consumers, using features offered by the control and forwarding planes. Traditional service guarantees are associated with

enablers that will support operators to "unlock" the 5G-related enterprise opportunities[102],[103]. Network slicing can enable operators to create pre-defined, differing levels of services to different enterprise verticals, making them able to customise their own operations. Currently, operators and their vendors are working to make the cost of deploying network slicing "marginal" to the broader investment case for 5G. This is an important consideration because building a new investment case to retrofit an already deployed network is difficult and slow[104]. A direct benefit to operators is the capacity to deploy only the functions needed to support specific customers and specific market segments[105]. Additional functionality not needed for the particular customer or market segment need not be deployed. This implicates for significant benefits in savings in investments and operational expenses. Furthermore, this supports faster deployment of 5G systems, as only dedicated parts of the necessary functionalities will be deployed, enabling faster time-to-market.

4.1.4 Discussion upon the network slicing-related conceptual approach

Figure 44, *below*, illustrates the concept of network slicing[106]. A network slice is composed of a collection of the necessary network functions (e.g. virtualised and/or physical, shared or slice specific) that support the service requirements of particular usage scenario(s). It is required to be possible to associate devices to their proper slices in order to "meet" the respective requirements, based on device or service type. For example, a mobile phone requesting high definition (HD) streaming service is associated to an eMBB (enhanced Mobile Broadband) slice. While the mMTC (massive Machine Type Communications) slice and the URLLC (Ultra-Reliable Low Latency Communications) slice are respectively allocated to a sensor measuring the temperature for an agricultural application and an automobile using V2X[107],[108] (Vehicle-to-Everything) communication

---

resource attributes such as throughput, packet loss, latency, network bandwidth/burst or other bit rates and security. In addition, redundancy and reliability are provided by the infrastructure to improve overall QoE. More recently, concepts such as edge computing allow opportunistic placement of services to meet stringent requirements of low latency and/or high bandwidth applications.

[101] For mobile operators to be able to slice the network assets dynamically, they will have to rely upon analytics. Network analytics will be able to model traffic patterns for operators so they can predict patterns. Thus, analytics is about understanding and predicting how one or more resources might be used.

[102] In the 5G era, different industry verticals are seeking to leverage the power of technology to boost productivity across swathes of the economy. Network slicing builds on this expectation, and together with the promise of Massive IoT and ultrareliable/low latency services, can effectively support the transformation of vertical industries.

[103] T. Yoo (2016): "Network Slicing Architecture for 5G Network". In IEEE (Ed.), Proceedings of the 2016 International Conference on Information and Communication Technology Convergence (ICTC-2016). Jeju, South Korea, October 19-21, 2016.

[104] That is the lesson learned from the past where VoLTE launches were decoupled from LTE launches (compared to 2G/3G voice services).

[105] One example of use case for dynamic network slicing is in a stadium where network traffic escalates before the game starts but it is primarily on the downlink because event-goers are downloading maps or getting information about the event that is happening in the stadium. Once the game -or event- begins, the traffic shifts to the uplink because people are uploading video and photos to social media. The challenge for the operator is to "tailor" the network based upon the traffic pattern in the respective operational environment.
Also see, *for example*: A. Kostopoulos, I.P. Chochliouros, E. Sfakianakis, D. Munaretto and C. Keuker (2019): "A Cloud-based Architecture for Video Services in Crowd Events". In Springer Nature Switzerland AG (Ed.), Proceedings of the 15th International Conference on Artificial Intelligence Applications and Innovations (AIAI-2019) - 4th Workshop on 5G - Putting Intelligence to the Network Edge (5G-PINE), pp. 7-18. Hersonissos, Crete, Greece, May 24-26, 2019.

[106] International Telecommunication Union - Telecommunications Standardization Sector (ITU-T) (2008, December): ITU-T Recommendation Y.3112 (12/2008): "Framework for the Support of Multiple Network Slicing". Available at: *https://www.itu.int/rec/T-REC-Y.3112-201812-I/en*

[107] For this specific case also see, *among others*: M.-P. Odini (2017, December):"V2X and Network Slicing", *IEEE Softwarization Magazine*. Available at: *https://sdn.ieee.org/newsletter/december-2017/v2x-and-network-slicing*

[108] C. Campolo, A. Molinaro, A. Iera, and F. Menichella (2017, December): 5G Network Slicing for Vehicle-to-Everything Services. *IEEE Wireless Communications Magazine*, vol. 24, no. 6, pp. 38-45.

service[109]. The network slicing concept allows the respective 5G network operator to provide dedicated logical networks (i.e. network slices) with customer specific functionalities. A network slice, spanning all the network segments including radio access network, transport network and core network, can be dedicated to specific types of service. When a User Equipment (UE) is only associated to a single dedicated network slice, the respective network can identify the association of the UE with the network slice based on user subscription, context, service provider's policy, etc. Otherwise, if a UE accesses multiple network slice instances simultaneously, it is recommended that the UE provide information to the network to assist the network slice selection process. For each network slice, dedicated resources (e.g. virtualised network functions, network bandwidth, QoS) are allocated and an error or fault that occurs in one slice does not cause any effect in other slices.

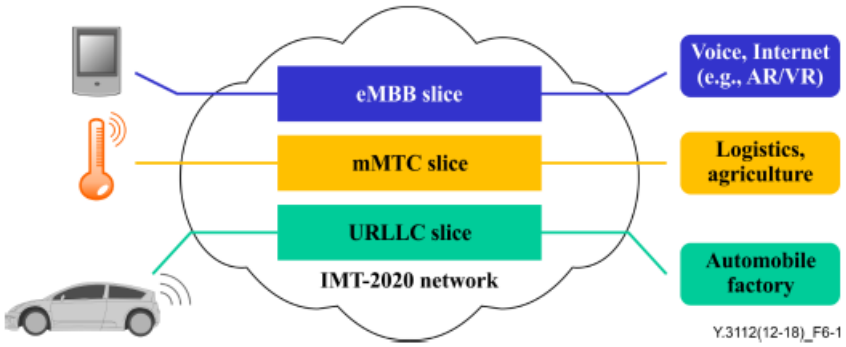


Figure 44: The concept of network slicing [Source: ITU-T Recommendation Y.3112 (12/2008)].

In mobile networks of previous generations, mainly serving mobile phones, these have been optimized for voice and Internet services in a centralised deployment. However, in the modern 5G network, they have to serve a variety of devices with different requirements and different needs. The above-mentioned usage scenarios require different requirements on performance. For instance, the mMTC usage scenario that connects stationary sensors measuring temperature, humidity, precipitation, etc. to mobile networks does not require features like handover or location update, which have been critical in serving mobile phones. For the eMBB usage scenario, high data rates are driven by the increasing use of data for services such as streaming (e.g. 4K/8K UHD (Ultra-High Definition)) and interactive services (e.g. augmented reality (AR)). These services come with stringent requirements for user experienced data rates as well as associated latency requirements to meet service requirements[110]. On the other hand, the URLLC usage scenario (e.g. autonomous driving, remote controlled robots) requires, unlike mobile broadband services, a substantially lower latency as well as higher reliability.

Regarding a possible network architecture-related approach, as depicted in Figure 45, the network slicing concept actually consists of 3 layers: (i) Service Instance Layer; (ii) Network Slice Instance Layer, and; (iii) Resource layer.

The Service Instance Layer represents the services (end-user service or business services) which are to be supported. Each service is represented by a Service Instance. Typically, services can be provided

[109] An interesting context has also been proposed in: T. Soenen, R. Banerjee, W. Tavernier, D. Colle, and M. Pickavet (2017): "Demystifying Network Slicing: From Theory to Practice". In IEEE (Ed.), Proceedings of the 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM). Lisbon, Portugal, May 08-12, 2017.

[110] International Telecommunication Union - Radiocommunications Sector (ITU-R) (2017, November): Report ITU-R M.2410-0 (11/2017): "Minimum requirements related to technical performance for IMT-2020 radio interface(s)". Available at: *https://www.itu.int/pub/R-REP-M.2410-2017*

by the network operator or by third parties. In line with this, a Service Instance can either represent an operator service or a 3rd party provided service. A network operator uses a Network Slice Blueprint to create a Network Slice Instance. A Network Slice Instance provides the network characteristics which are required by a Service Instance. A Network Slice Instance may also be shared across multiple Service Instances provided by the network operator. The Network Slice Instance may be composed by none, one or more Sub-network Instances, which may be shared by another Network Slice Instance. Similarly, the Sub-network Blueprint is used to create a Sub-network Instance to form a set of Network Functions, which run on the physical/logical resources.

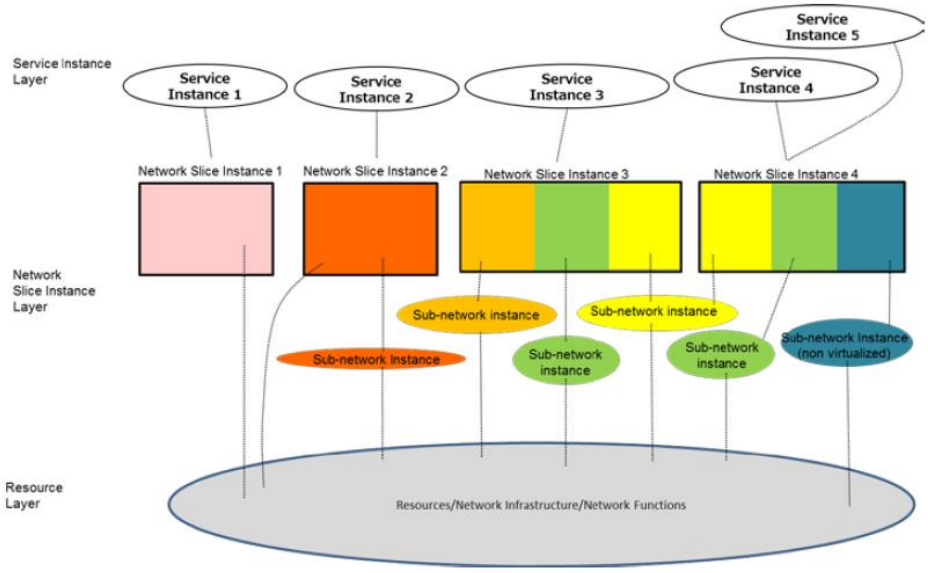


Figure 45: Network slicing conceptual outline.

Definitions are given as follows:

- *Service Instance:* An instance of an end-user service or a business service that is realised within or by a Network Slice
- *Network Slice Instance:* a set of network functions, and resources to run these network functions, forming a complete instantiated logical network to meet certain network characteristics required by the Service Instance(s).
  - A network slice instance may be fully or partly, logically and/or physically, isolated from another network slice instance.
  - The resources comprises of physical and logical resources.
  - A Network Slice Instance may be composed of Sub-network Instances, which as a special case may be shared by multiple network slice instances. The Network Slice Instance is defined by a Network Slice Blueprint.
  - Instance-specific policies and configurations are required when creating a Network Slice Instance.
  - Network characteristics examples are ultra-low-latency, ultra-reliability, etc.
- *Network Slice Blueprint:* A complete description of the structure, configuration and the plans/workflows for how to instantiate and control the Network Slice Instance during its life cycle. A Network Slice Blueprint enables the instantiation of a Network Slice, which provides certain network characteristics (e.g.: ultra-low latency, ultra-reliability, value-added services for enterprises, etc.). A Network Slice Blueprint refers to required physical and logical resources and/or to Sub-network Blueprint(s).

- *Sub-network Instance:* A Sub-network Instance comprises of a set of Network Functions (NFs) and the resources for these NFs.
  - The Sub-network Instance is defined by a Sub-network Blueprint.
  - A Sub-network Instance is not required to form a complete logical network.
  - A Sub-network Instance may be shared by two or more Network Slices.
  - The resources comprise of physical and logical resources.
- *Sub-network Blueprint:* A description of the structure (and contained components) and configuration of the Sub-network Instances and the plans/workflows for how to instantiate it. A Sub-network Blueprint refers to Physical and logical resources and may refer to other Sub-network Blueprints.
- *Physical resource:* A physical asset for computation, storage or transport including radio access
  - Network Functions are not regarded as Resources.
- *Logical Resource:* Partition of a physical resource or grouping of multiple physical resources dedicated to a Network Function or shared between a set of Network Functions.
- *Network Function (NF):* Network Function refers to processing functions in a network.
  - This includes but is not limited to telecom nodes functionality, as well as switching functions e.g. Ethernet switching function, IP routing functions.
  - VNF is a virtualised version of a NF (refer to ETSI NFV for further details on VNF).

From the operational perspective and based on the 3GPP requirements, network operators shall have the following basic requirements for network slicing operations: (i) Create and manage network slices instances; (ii) create and manage services to slices pairing function; (iii) create network slices blueprints; (iv) deploy and operate fault, configuration, accounting, performance and security (FCAPS) functions for the defined network slices. Configuration management (CM), fault management (FM) and performance management (PM) capabilities should be provided per slice so operators can monitor end-to-end service health, including relevant information about the slice's performance relative to the required QoS from the slices, and also about the NFs' performance; (v) create and manage slice components pairing function (e.g. CN slice to RAN slice pairing rules); (vi) modification of network slices will be possible, such as adding, deleting and modifying network slices; (vii) slices are end-to-end[111] (e.g. RAN, CN),; (viii) resource management for a network slice may cross operator domains, and so may require cooperating resource management domains, and; (ix) slice as-a-service concept allows a third party to create and manage a network slice configuration (e.g. scale slices) via suitable APIs, within the limits set by the operator.

Today's networks were designed for the delivery of personal communication services and content, such as voice, video, and web browsing to a variety of devices. But 5G networks must enable and support a fully mobile and connected society. 5G will qualify new verticals, new services and new business models that are not possible or hands-on with 4G and other legacy mobile technologies.

[111] I. Afolabi, M. Bagaa, T. Taleb, and H. Flinck (2017): "End-to-end network slicing enabled through network function virtualization". In IEEE (Ed.), Proceedings of the 2017 IEEE Conference on Standards for Communications and Networking (CSCN-2017). Helsinki, Finland, September 18-20, 2017.

With 5G, a proliferation of wearable devices and wireless connected objects will pave the way to a wide variety of new uses[112]. More pervasive human-centric applications will be launched, such as wearables for advanced telemedicine applications, virtual/augmented reality and UHD video. Simultaneously, machine-to-machine (M2M) and human-machine type applications can be developed, aiming to enhance the standards of our live, especially by promoting safety and innovation. Such potential applications can promote growth in a great variety of (vertical) industries, including public safety, health, logistics, smart city, manufacturing, energy, media, and automotive[113]. Beyond connectivity, 5G offers unique opportunities to operators to create new business models for consumers, enterprises, and industry specific services, as well as content and application providers.

4.1.5 Network slicing status

The idea of slicing a common IT infrastructure into fully-fledged isolated virtual test-bed environments for "planetary-scale service development" was introduced in 2002, in the context of PlanetLab[114]. Recently, the concept has been described by NGMN in the context of telecom networks [22] [20]. Currently, there are several definitions of network slicing, but the NGMN description is widely used (for example by 3GPP). It is worth mentioning that the NGMN concept treats services as external to slices and allows for concatenation of slices or sub-slices. The NGMN specification lacks some basic functions necessary for network slicing, like slice description, slice selection (matching) and runtime, as well as life-cycle management of slices.

4.1.6 Standardization of network slicing

Several Standards Developing Organizations (SDOs, cf. International Telecommunication Union[115] (ITU), 3rd Generation Partnership Project[116] (3GPP), Internet Engineering Task Force[117] (IETF)) are currently active in the standardization of network slicing. None of these activities has led to a definition of the complete solution, yet. In fact, the standardization of network slicing concerns so far selected mechanisms only. An overview of network slicing standardization efforts can be found in [23].

The 3GPP defines network slices for mobile networks, especially in order to handle diverse requirements of 5G service classes, i.e. enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC) and massive Machine Type Communications (mMTC). In 5G Release 15 series of 3GPP specifications, the 5G Core Network (5GC) has some functional blocks responsible for network slicing. There are some functions of the control plane of 5GC as Access and Mobility Management Functions (AMF), which have been defined as common to multiple (or all) slices. In [24] the 3GPP has defined a Single Network Selection Assistance Information (S-NSSAI) that includes information about the slice/service type (SST) and an additional parameter called Slice

[112] For more informative details about potential use cases also see, *inter-alia*: Alliance for Telecommunications Industry Solutions (ATIS) (2015, November): "5G Reimagined: A North American Perspective", White Paper. Available at: *https://access.atis.org/apps/group_public/download.php/27373/ATIS-I-0000050.pdf*

[113] With the introduction of slicing, operators have the opportunity to differentiate through service customization. In the case of connected cars, *for example*, different car vendors may be operated in different slices. Operators get the ability to allocate dedicated virtual resources to different car manufacturers and operate them independently. Each car manufacturer may be equipped with different services such as traffic efficiency, traffic safety, infotainment, security and emergency support. Service differentiation allows mobile operators to create different pricing strategies.

[114] "A slice is a set of allocated resources distributed across PlanetLab. To most users, a slice means UNIX shell access to a number of PlanetLab nodes[...] Sliver: A set of allocated resources on a single PlanetLab node". Definition in PlanetLab main website: *https://www.planet-lab.org/doc/guides/user* [Accessed April 2019].

[115] ITU main website: *https://www.itu.int/en/Pages/default.aspx* [Accessed April 2019].

[116] 3GPP main website: *https://www.3gpp. org/* [Accessed April 2019].

[117] IETF main website: *https://www.ietf.org/* [Accessed April 2019].

Differentiator (SD). The 3GPP terminal is allowed to request attachment to up to eight network slices at the same time. The 3GPP is also working on standards related to slice life-cycle: provisioning, fault, configuration and performance management [25] [26] [27]. The slice-related high-level management architecture has been proposed in the 3GPP study, but it still remains absent in the 3GPP management architecture [28] [29]. The 3GPP has included network slicing in Release 15 of its specifications and more network slicing mechanisms will be included in Release 16 that is expected to be published by March 2020. So far, the 3GPP vision concerns only a single operator of the network and slices, as well as the infrastructure. Therefore, no business interfaces have been defined.

There are some network slicing activities at ITU-T, Study Group 13[118] (SG-13), especially within Questions 20 and 21. The term "network slice" (a complete logical network to provide certain network characteristics) has been initially defined in the context of the architecture of mobile core networks [30]. Later, the approach has been generalised to the entire communication network architecture. The ITU-T has already published a series of documents and recommendations concerning network softwarization and slicing framework [31], slice life-cycle management and orchestration, the high-level architecture of the slice management plane [32] and mechanism for multiple slice attachment of user equipment (UE) [33]. The recommendation on network slicing business models is consented and is about to be published soon[119]. There are on-going works on the architecture of IMT-2020[120] network, its capabilities exposure and network slice orchestration and management. This work is performed as part of a dedicated Focus Group, although they are currently in draft form (Table 3).

Table 3: IMT-2020 Focus Group reports.

Draft reports from IMT-2020
Draft Terms and definitions for IMT-2020 in ITU-T (O-040)
Draft Technical Report: Application of network softwarization to IMT-2020 (O-041)
Draft Recommendation: Requirements of IMT-2020 from network perspective (O-042)
Draft Recommendation: Framework for IMT-2020 network architecture (O-043)
Draft Recommendation: Requirements of IMT-2020 fixed mobile convergence (O-044)
Draft Technical Report: Unified network integrated cloud for fixed-mobile convergence (O-045)
Draft Recommendation: IMT-2020 network management requirements (O-046)
Draft Recommendation: Network management framework for IMT-2020 (O-047)

There are some network slicing-related Birds of a Feather[121] (BOF)-level activities in IETF, but no IETF Working Group (WG) is so far dedicated to network slicing. At IETF the network slicing is discussed in terms of terminology, problem statement, cross-domain architecture and use cases [23]. A network slice information model has been proposed.

Although the ETSI NFV concept [21] and the definition of MANO architecture [34] are fundamental for the network slicing technology, the issue of slicing was absent in the first two releases of the NFV program. The currently on-going Release 3 upgrades the standards with the matter; the recommendations for necessary updates of NFV normative documents are already published [35], but the works on full adoption of the recommendations are still in progress.

[118] ITU-T Study Group 13: *https://www.itu.int/net4/ITU-T/lists/loqr.aspx?Group=13&Period=16* [Accessed April 2019].

[119] ITU-T, "Business Role-based Models in IMT-2020", Draft recommendation ITU-T Y.3103. Source: *https://www.itu.int/ITU-T/workprog/wp item.aspx?isn=14505* [Accessed April 2019].

[120] ITU-T Focus Group on IMT-2020: *https://www.itu.int/en/ITU-T/focusgroups/imt-2020/Pages/default.aspx* [Access April 2019].

[121] Birds of a Feather meetings: *https://tools.ietf.org/bof/trac/* [Accessed April 2019].

The GSMA[122] analyses the network slicing from the point of view of business customers [36]. In GSMA, network slicing is understood as virtually independent business operations on a common physical infrastructure implemented in an efficient and economical way. The defined by GSMA slice features include data speed, latency, reliability, security and slice services. In the concept, a network slice can be used by several operators.

The overview of network slicing standardization by different organizations presented above shows that different SDOs are focused on different aspects of network slicing, but no unified, single definition and approach exist so far.

4.1.7 Network slicing-related research activities

In the European H2020-5G PPP Research Program[123], there are many projects related to network slicing. Each of them tries to address different aspects of network slicing.

The **5G NORMA**[124] project developed the architecture of multi-tenant Software-Defined Mobile Network (SDMN). The network is split into common (inter-slice) and dedicated (intra-slice) functional parts. The abstracted management interface of functions comprising the control and data planes is provided by the inter-slice SDMN-Coordinator and the intra-slice SDMN-Controller, which interact with the ETSI NFV MANO layer extended with Element/Network Managers and the Inter-slice Resource Broker – the master entity, able to expose mobile network slice service to higher layers.

The **5G Exchange**[125] (5GEx) project is focused on a multi-provider orchestration framework for transport networks. Part of the framework is a business layer for multi-provider collaboration. The framework enables capability and topology information exchange between the operators, a collection of service KPIs and provides support for life-cycle management of deployed services.

The **SELFNET** project[126] worked on providing an autonomic and cognitive (AI/ML-based) management framework for softwarized multi-tenant environments and supporting self-healing, self-protection, self-optimization and automated deployment.

The **SESAME**[127] project proposes the RAN environment based on cloud-enabled small cell concept which supports virtualisation and partitioning of small cells into logically isolated and individually profiled resource slices utilised by multiple tenants. The key problems addressed by 5G SESAME are "self-X" features, radio access management techniques, orchestration, NFV management and management by a tenant.

The **5G ESSENCE**[128] project develops an architecture that exposes a shared RAN infrastructure (including virtualised execution environment) to multiple tenants, providing end-to-end slicing mechanisms with customization on a per-tenant basis. It proposes the way of integration of the traditional network management with NFV operations.

---

[122] Global System for Mobile Communications (GSM) Association: *https://www.gsma.com/* [Accessed April 2019].

[123] 5G research in H2020: *https://5g-ppp. eu/european-5g-actions/* [Accessed April 2019].

[124] 5G NORMA ("5G Novel Radio Multiservice adaptive network Architecture") project (GA No. 671584). Deliverable 3.3: "5G NORMA network architecture final report", October 2017.

[125] 5GEx ("5G Exchange") project (GA No. 671636): *https://5g-ppp. eu/5gex/* [Accessed April 2019].

[126] Selfnet ("A Framework for Self-Organised Network Management in Virtualised and Software Defined Networks") project, (GA No. 671672), *https://selfnet-5g.eu/* [Accessed April 2019].

[127] SESAME ("Small CEllS coordination for Multi-tenancy and Edge services") project, (GA No. 671596). Deliverable 2.5: "SESAME Final Architecture and PoC Assessment KPIs", December 2106.

[128] 5G ESSENCE ("Embedded Network Services for 5G Experiences") project (GA No. 761592): *http://www.5g-essence-h2020.eu/* [Accessed April 2019].

The **5G-TRANSFORMER**[129] project proposes an SDN/NFV-based 5G mobile transport network relying on network slicing and multi-access edge computing (MEC) customized for vertical industries. The proposed architecture offers open APIs to the verticals that are used for the definition of the vertical services and associated Service Level Agreements (SLAs). The east-westbound interfaces, of the approach, enable service and resource federation among different administrative domains by peering with other providers.

The **5G-MoNArch**[130] project is dedicated to extensions of 5G architecture related to inter-slice context-aware optimization (inter-slice context sharing, coordination and adaptive slice selection), and inter-slice management & orchestration (slice admission control and cross-slice congestion control). The project is focused on methods and algorithms for optimization of orchestration across network slices and among virtualised functions.

The **SliceNet**[131] project addresses the issues of the data plane, control plane and management plane in network slicing across multiple administrative domains. In the context of the management plane, it plans to integrate the slice FCAPS functionalities and use a cognition-based QoE management approach. For the orchestration plane, SliceNet proposes a cross-plane coordination approach and will design a cross-plane orchestrator that preserves the boundaries between the planes.

The **MATILDA**[132] project plans to develop intelligent, unified and hierarchical, application level orchestration mechanisms for the automated placement of the 5G-ready applications and the creation and maintenance of their network slice instances. The project will use the concept of "intent" to request, negotiate, deploy, maintain and terminate slice instances customized for a specific application. The slice intent will be automatically or semi-automatically translated by the orchestrator into low-level primitives. The project will focus on the entire life-cycle of the 5G-ready applications, in terms of development, deployment and operations phases.

The **NECOS**[133] project introduces the "Service Broker" – the entity exposing the consolidated catalogue of 3rd party services (Infrastructure as-a-Service (IaaS), Platform as-a-Service (PaaS), Software as-a-Service (SaaS), VNF as-a-Service (VNFaaS), Security as-a-Service (SECaaS) and Management as-a-Service (MaaS)).

The **5G!Pagoda**[134] project addresses slice management scalability and proposes the in-slice management paradigm combined with autonomic slice management. Each 5G!Pagoda compliant slice consist of slice "core functions", as well as embedded slice management and slice operations support parts.

The 5G PPP has published an integrated result of some of the 5GPP H2020 projects as architecture specification that includes the network slicing [37]. The mentioned document can be however seen as a high-level working draft only. Most of the mentioned projects aim to demonstrate their results as a "proof of concept (PoC)", however, so far none of them has demonstrated a carrier-grade solution.

---

[129] 5G Transformer ("5G Mobile Transport Platform for Verticals") project, (GA No. 761536): *http://5g-transformer.eu/* [Accessed April 2019].

[130] 5G-MoNArch ("5G Mobile Network Architecture for diverse services, use cases and application in 5G and beyond") project (GA No. 761445). Deliverable 2.2: "Initial overall architecture and concepts for enabling innovations", July 2018.

[131] SLICENET ("End-to-End Cognitive Network Slicing and Slice Management Framework in Virtualised Multi-Domain, Multi-Tenant 5G Networks") project, (GA No. 761913), *https://slicenet.eu/* [Accessed April 2019].

[132] MATILDA ("A holistic, innovative framework for the design, development and orchestration of 5G-ready applications and network services over sliced programmable infrastructure) project, (GA No. 761898*), http://www.matilda-5g.eu/* [Accessed April 2019].

[133] NECOS ("Novel Enablers for Cloud Slicing") project (GA No. 777067). Deliverable 2.1: "Initial definition of use cases", Jun. 2018. Available at: *http://www.h2020-necos.eu/documents/deliverables/*

[134] 5G!Pagoda project (GA No. 723172): *https://5g-pagoda.aalto.fi/* [Accessed April 2019].

4.2 Open issues

This section collects the network slicing-related issues that have to be solved before the technology can be deployed in commercial, large-scale networks. Some of them are already partly addressed, but some of them, mostly business-oriented (business interfaces, accounting, lawful intercept), still require intensive research.

4.2.1 Addressing highly differentiated slicing needs and contexts

One of the most important problems related to network slicing is the differentiation of slices in terms of functionalities, geographic span, longevity, and management- and orchestration-related requirements. The mentioned differences make a uniform approach to network slicing troublesome:

Some slices can be implemented by a single entity that owns the infrastructure, slice orchestrator and operates (manages) its slices. However, in some cases, it may be needed to use the infrastructure owned by multiple infrastructure providers and allow the slice tenants to manage and even to orchestrate their slices. The multi-provider and the multi-tenant cases are much complicated as they require additional interfaces between 3rd parties. All these interfaces are not necessary in the previous case.

Some slice functional requirements may lead to the involvement of certain hardware-based solutions (e.g. RAN nodes) that may limit slice deployment in a “visited” domain. All such limitations have to be analysed in advance.

Some slices may be deployed in a relatively small area (e.g. stadium, factory) whereas the other ones may require national or international span. The “wide area” slices require additional mechanisms to cope with scalability and performance issues related to the potentially huge number of involved nodes.

Some slices may have a relatively limited number of users (e.g. videoconferencing) whereas the other ones may have a huge number of users (e.g. video streaming from an event). The latter case raises important scalability issues which do not exist in the previous case.

Some slices can be almost static, and their deployment time is not critical one (e.g. the Mobile Virtual Network Operator (MVNO) case), but other slices can stay for a relatively short period of time and their deployment time can be critical. The second case concerns on-demand created slices; *for example*, for videoconferencing or an emergency. Such cases can be supported by special mechanisms that will contribute to the minimization of the slice deployment time. They can include the creation of some functions that are shared by multiple slices and/or deploying a priori certain slices and keeping them in a “frozen” state by minimization of resources that are used by them.

Allocation of resources to slices should be prioritized. As all the slices share common infrastructure appropriate policy of resource to slice allocation must be implemented. There will be some slices which should have high priority of resource allocation. It especially concerns the slices that are used for handling emergency situations, but it can also be driven by the business goals.

All the above-mentioned issues show that the architecture and solution able to cope with all of them could be very complex and include certain mechanisms that in some cases are not needed. A solution to the problem is the definition of profiles or variants of the network slicing architecture customized to the use cases. So far, no such activities have been noticed.

4.2.2 Architecture-related issues

Despite many network slicing approaches, there is so far no approach that considers all carrier-grade operator needs. The 3GPP approach is oriented towards mobile networks, ignores transport network slicing, multi-domain slicing and does not cope with multi-tenancy or multi-provider infrastructure.

The RAN slicing is not defined in details and the way in which RAN is sliced is left to the implementers.

In fact, the operators need an architecture that can cope not only with mobile networks but also with other technological solutions. Moreover, the approach should reuse the already installed, legacy equipment and provide a migration path from the "old style" networking to the virtualised one. The issues of network slicing efficiency, performance monitoring and control, security, accounting and global interoperability must be a part of the overall picture. As it has been noticed previously, different slice use cases may require different approaches to the orchestration, deployment and business-oriented interfaces. It seems it is possible to develop an architecture for each use case, but it should have a part that is common for all the use cases and enhanced for each of the use cases. In order to achieve such goal, a top-down approach with a collection of requirements for all the use cases is needed.

4.2.3 Slice description and selection issues

An important feature of a slice is slice descriptor that is used for slice selection or concatenation of sub-slices in order to create an end-to-end slice. A step in this direction is provided by 3GPP, but more work on the topic is needed that probably will lead to the definition of slice descriptors which will be used by globally available services, like *Netflix*[135] or *YouTube*[136].

4.2.4 Service to slice relationship

In the NGMN concept, the slice services are external to a slice, although such a slice has been created to support a specific service or set of services. In practice, some slices will have an embedded service platform (i.e. single-service slices) whereas for other types of slices will support multiple services which have a different life-cycle than the slice supporting them. In the latter case, the slice should expose an interface to services (often called Slice Exposure Function). Moreover, new services should be able to reuse the already deployed services. For that purpose, it is possible to use the vertical stitching of slices paradigm, in which services are also seen as slices (the service platforms for their deployment similarly to "network slices" needs an allocation of computing, connectivity and storage resources).

4.2.5 Data plane issues

The data plane in network slicing should handle all slices efficiently, providing isolation between them as well as differentiated QoS. Some slices should provide minimal end-to-end transmission delay (in the range of single milliseconds) that raises serious issues about the usage of virtualised data plane and using virtual networking devices. As it has been noticed, the communications between VMs[137] and containers[138] introduces a delay that cannot be ignored. Using some additional software-or hardware-related data plane accelerators limits the flexibility of slice deployment; however, in some cases it can be hard to avoid such solutions. The QoS of the data plane of the sliced network is an extended version of the classical QoS, often linked with the Differentiated Services

---

[135] *Netflix* is a global-scale streaming service, main website: *https://www.netflix.com/* [Accessed April 2019].

[136] *YouTube* is a video hosting platform, main website: *https://www.youtube.com/* [Accessed April 2019].

[137] In computing, a virtual machine (VM) is an emulation of a computer system. Virtual machines are based on computer architectures and provide functionality of a physical computer. Their implementations may involve specialised hardware, software, or a combination. For more informative details also see, *inter-alia*: *https://en.wikipedia.org/wiki/Virtual_machine*

[138] For more details see, *for example*: *https://en.wikipedia.org/wiki/Operating-system-level_virtualization#cite_note-1*

(DiffServ[139]) or Integrated Services (IntServ[140]) architecture and should include also different levels of security and reliability (e.g. for URLLC). Some of the parameters should be slice-specific, but within the slice, there can be also data streams that may require differentiated handling. There is an open issue, whether the reliability will be either provided by the slice infrastructure or it should be implemented as embedded slice mechanisms, like the multi-path transmission, etc. In practice, probably both mechanisms will have to cooperate. There is no research activity on this topic, so far.

4.2.6 RAN issues

The RAN slicing brings a lot of problems. The Software-Defined Radio (SDR) concept (already in use by the operators of 2G/3G/4G networks in Common Public Radio Interface (CPRI)-based solutions) allows for relatively simple network slicing by building completely isolated networking solution. However, the effectiveness of such solution(s) can be questioned as each of the RAN slices will replicate all the control channels of the mobile network. It is therefore also possible and desirable to allow for RAN slices to share control channels and physical operations but, *at the same time*, separate the data plane of each slice and provide to each slice mechanisms for customizing the control plane (for example, paging intervals, radio access bearers set-up or handover handling according to slice-specific requirements). The mentioned approach means that some of the control plane functions have to be common to several or all slices. The radio resource allocation to a slice is another issue that has to be handled efficiently. Moreover, incorporating multi-RAT slicing, in which the terminal attached to a slice can change the radio technology (a mechanism that in present networks is called a vertical handover), requires additional mechanisms related to radio resources allocation. The proposed, new split of RAN functions into RRU, CU, DU makes the implementation of RAN slicing even harder. On the other hand such split gives new possibility of creating RAN partitions that may be included the 'slice-owned' RAN functions. The case provides better isolation and easier guarantee of QoS parameters (like data plane delay) at the price of less efficient usage of resources (limited aggregation gain).

4.2.7 Slice isolation issues

One of the "key" features of the network slicing approach is isolation between slices. Such isolation can be done in different ways. The basic assumption is that the allocation of virtual resources (including connectivity) provides slice isolation already. As it has been mentioned in case of RAN slicing, some control plane functions can be common for multiple slices what reduces the inter-slice isolation. In RAN the data plane isolation can be achieved by ciphering of radio access bearers (as it is already done in deployed 3GPP networks). The isolation of slices means also the separation of slice-oriented operations including slice management and orchestration. Both types of operations should have isolated partitions in the overall orchestration and management system, and access to them has to be secure.

---

[139] "Differentiated Services is a multiple service model that can satisfy differing QoS requirements. With Differentiated Services, the network tries to deliver a particular kind of service based on the QoS specified by each packet." Definition by Cisco: *https://www.cisco.com/c/en/us/td/docs/ios/12_2/qos/configuration/guide/fqos_c/qcfdfsrv.html* [Accessed April 2019].

[140] "Cisco IOS Software supports two fundamental Quality of Service architectures: Differentiated Services (DiffServ) and Integrated Services (IntServ). In the DiffServ model a packet's *class* can be marked directly in the packet, which contrasts with the IntServ model where a signaling protocol is required to tell the routers which flows of packets requires special QoS treatment. DiffServ achieves better QoS scalability, while IntServ provides a tighter QoS mechanism for real-time traffic. These approaches can be complimentary and are not mutually exclusive." Definition by Cisco: *https://www.cisco.com/c/en/us/products/ios-nx-os-software/integrated-services/index.html* [Accessed April 2019].

4.2.8 Management issues

The network slices can be seen as parallel networks (some-times combined with services) that share a common infrastructure. The average number of slices in a telco operated infrastructure is not known yet, but it is expected to be huge. It raises the problem of the management of such a solution as the network management is well-known to be complex even in a single network. The management issues will grow not only with the number of slices but also depends on their complexity. The human-operated management is not applicable in such a case, and a kind of autonomic or cognitive management should be implemented to cope with management scalability. In some cases, the delegation of the slice management to slice tenant can be required. This type of management should be relatively simple, comfortable for tenants and use the intent management paradigm. Some short-lived slice should be self-managed. A concept of in-slice management that delegates some management functions to slice tenants is presented in [38].

4.2.9 Orchestration issues

The slice orchestrator is in general responsible for slice life-cycle management, as well as slice performance management and faults handling. The ETSI NFV MANO framework [34] is used in many network slicing approaches for the implementation of the network slicing. Most of them use a single orchestrator that raises issues related to orchestrations scalability and performance. The use of multiple orchestrators raises their own problems related to the information exchange between the orchestrators that can be solved in a peer-to-peer (P2P) or hierarchical way. So far, the hierarchical orchestration is not properly supported by ETSI MANO. However, the “umbrella orchestrator” has been already introduced. It has to be noted, that the MANO framework has not been designed for the multi-provider and multi-tenant environment and therefore has no accounting interfaces and is not ready for multiple infra-structure providers. Moreover, the role of the OSS/BSS is still not clearly defined by ETSI, as well as the management of the orchestrator (including orchestration performance monitoring) is not included in ETSI specifications. At present, the ETSI MANO-compliant orchestrator has no mechanisms to cope with orchestration scalability issues.

Another orchestration problem is linked with the business model of network slicing. In most of the present approaches, the slice provider is the entity that owns the infrastructure, MANO orchestrator, OSS/BSS as well as slice templates and VNFs. It is, *however*, possible to give the orchestration capabilities to 3rd parties, so they may create their own slices. In such case, the involved 3rd party should have a dedicated orchestrator with the appropriate back-office. The existence of several orchestrators that use the same infrastructure requires some modifications of the MANO architecture. In the 3GPP Release 15 specifications the orchestration of slices is done by the MNO and some slice management capabilities can be delegated to slice tenant via the publish-subscribe mechanism.

4.2.10 Multi-domain slicing-related issues

An important value of the slicing concept is an end-to-end network slicing that crosses multiple technological and administrative domains. The existence of multiple domains raises a lot of issues related to the end-to-end (E2E) management and orchestration of slices. Deployment of a slice in multiple administrative domains requires proper mechanisms for negotiations between domains’ operators and accounting. Moreover, the implementation of the multi-domain slices can be performed in several ways. One of them lies on the orchestration of all the domains by a single operator controlling the resources of other domains; another possibility is to involve all the domains’ orchestrators of which each obtains from the operator that is the slice initiator its part of the slice template. Yet another way of multi-domain orchestration can be based on the domain-level

implementation of well-known slices that may be parametrized only by the slice initiator. Probably all the mentioned scenarios will have its application in real networks.

4.2.11 Business models of network slicing

There is no doubt that the network slicing paradigm is a big disruption of the business models of telco operators. In the future, there will be probably no physical network operators, but all the networking solutions (except the infrastructure) will be implemented as virtual ones. The business landscape will change significantly, and it will include infrastructure providers, slice providers (orchestrator operators), slice tenants, slice template providers, slice functions (VNFs) providers and probably some brokers. For each of them, appropriate interfaces of the functional architecture have to be provided in order to carry negotiations and accounting. This issue is so far almost neglected in all research projects and standardization efforts.

4.2.12 Network slicing and legal issues

The operators of wide area networks, including telcos, are obliged to provide mechanisms that allow (in legally justified cases) lawful intercept of voice, data transmission or SMSs. The same requirement will also concern network slices. However, the slicing will make the implementation of the legal intercept mechanism much harder than in the existing networks. It is caused by the immanent isolation of slices and the potential capability of the creation of in-slice mechanisms that will implement slice-specific encryption. It seems therefore that each slice template should provide an interface for legal intercept of the transmitted data or voice and before its deployment the mechanism should be validated by the regulator. In the case of international network slices several regulatory requirements probably will have to be addressed. The obligatory mechanism will add complexity to the overall network slicing architecture. So far, the legal intercept of network slices has been completely ignored by the scientific and standardization communities.

4.3 New KPIs for network slicing

Network slicing is still not a mature technology, and so far, the performance issues related to both the run-time and the life-cycle operations of network slicing are not “addressed” well. The set of the most important performance parameters of the telco system or subsystem is typically referred to as Key Performance Indicators (KPIs). In general, the set of KPIs can be huge, but using them and keeping conformance to them is a common practice of all telco operators. They use KPIs for (i) requirements for system definition and implementation; (ii) verification of proper functioning of installed networks, sub-networks, systems or sub-systems; (iii) definition of customer’s requirements and provider’s obligations as a part of the Service-Level Agreement (SLA) contract; (iv) comparison of performance provided by specific vendors or technologies. If a system is composed of multiple subsystems, fulfilling the subsystems’ KPIs can provide the fulfilment of the overall system’s KPIs. The definition of KPIs helps in the overall system engineering (dimensioning) in order to provide the requested end-to-end KPIs often linked with the Quality of Experience (QoE). For telco operators, verification of KPIs against their goals is a basic “health” indicator of solutions in service. Some KPIs can be standardized (as long as they are technology-specific), but some can be operator-specific. In general, standardization of KPIs is highly desirable, and Standards Developing Organizations such as ITU-T, ETSI, 3GPP or GSMA provide referential KPIs for some of their standardized solutions. In telco networks, KPIs may deal with different network segments, layers, mechanisms, aspects and also services or activities (e.g. fault handling time). Some of KPIs can be user-oriented, whereas others can be network-oriented. In mobile networks KPIs can be related to network transport, front-haul, radio link quality, data plane efficiency, and control plane operations as handover execution time, user

attachment time, etc. Typically, KPIs calculation is done by the network or service management system. However, some KPIs for the network management system itself (for example fault handling time) can be also defined. KPIs are usually calculated by the network/service monitoring system, and in some cases, the management system may have the ability to take actions in order to guarantee the requested KPIs. One of the issues related to KPIs is their efficient monitoring.

In the era of the softwarized network, there are new operations whose performance should be evaluated. For example, the orchestration, as well as virtual infrastructure performance, have to be monitored. The network slicing is perceived as a key technology that will be used by telco operators, soon. However, the KPIs concerning this technology remain still undefined. Such KPIs can deal with operations related to both, slice run-time operations as well as slice life-cycle operations. It is worth noting that the definition of KPIs is also linked with their calculation methodology and may have an impact on the overall architecture of the developed system. The following arguments are the main motivation for the paper that deals with KPIs for network slicing. This approach is linked with 5G (and 5G+) network slicing, but it can be used also in other networking solutions.

4.3.1 Background

The issue of network performance and resulting service quality is fundamental for telco operators. The numerous and specific technology-related performance indicators, when tracked, provide quantitative insight into the behaviour of equipment, sub-systems and entire systems. The group of higher-level abstraction performance indicators, giving a representative view of the end-to-end network, forms a list of Key Performance Indicators (KPIs), which at the level of underlying communication technology (network) contribute to the end-to-end communication service-level quality view, represented by Key Quality Indicators (KQIs). The referential approach to the topic is defined by several SDOs. It is worth emphasizing that the number of KPIs, in general, should be minimized, but at the same time, the KPIs should describe the most important system or network features. The currently used KQIs [39] [40] [41], offer a framework to assess objectively 2G/3G/4G services' performance and quality from the end-to-end perspective. The practical digest for operational use in telco operators, which integrates outputs of above-mentioned SDOs, is provided by GSMA [42] and its framework is structured in four layers:

- Network Availability;
- Network Accessibility;
- Generic service layer consisting of Service Accessibility, Service Integrity and Service Retainability;
- Specific service layer consisting of groups of parameters typical for distinct communication services (e.g. SMS, voice call, web browsing, streaming, etc.).

There is a lot of standardization efforts focused on defining of criteria of 5G network performance and quality assessment. The existing approaches are typically linked only with a partial view of the overall system. This is justified by the overall complexity of the 5G network and the way in which the 5G architecture is decomposed. The following table compares the current approaches (

Table 4).

Table 4: Comparison between Quality Assessment Approaches.

Quality Assessment Approach	Description	Sources
Based on service requirements and	Scenario-driven approaches that take into account the needs of current use cases or	"Minimum requirements related to technical performance for IMT-2020 radio interface(s)", ITU-R M. [IMT-2020.TECH PERF REQ] (draft), Feb. 2017.
		"Service requirements for next generation new

characteristics	market needs.	services and markets", 3GPP TS 22.261, ver. 16.6.0, Dec. 2018.
Based on next generation expectations	These focus on the requirements for 5G and next generation networks	"Recommendations for NGMN KPIs and Requirements for 5G", NGMN Alliance, Jun. 2016.
		"Study on scenarios and requirements for next generation access technologies", 3GPP TR 38.913, ver. 15.0.0, Jul. 2018.
Based on specific key technologies	Based on Management and Orchestration (MANO), providing Accessibility, Integrity and Utilisation KPIs and assessment of both MANO and network performance	"Management and orchestration; 5G end to end Key Performance Indicators (KPI)", 3GPP TS 28.554, ver. 15.1.0, Dec. 2018.
		"Management and orchestration; 5G performance measurements", 3GPP TS 28.552, ver. 16.0.0, Dec. 2018.
		"Network Functions Virtualisation (NFV) Release 2; Management and Orchestration; Performance Measurements Specification", ETSI GS NFV-IFA 027 V2.4.1, May 2018.

In the context of network slicing, worth mentioning are also standardization activities of ETSI NFV Industry Specifications Group[141], which has published several specifications dealing in general with performance monitoring within the NFV framework. A very detailed list of NFVI metrics has been developed by the Alliance for Telecommunications Industry Solutions Deployments (ATIS) [43]. However, both activities are focused on detailed performance monitoring for the internal MANO purpose and less linked with the KPIs philosophy.

There are also research projects that are focused on 5G network KPIs. The 5GENESIS[142] project, which is focused on 5G trials, plans to assess KPIs concerning: services (e.g. data rates, reliability, and latency), applications (user-perceived Quality of Experience – QoE and security), network (coverage and density) and network life-cycle management. Performance metrics such as slice establishment time, VNF relocation and instantiating times, and the computational resource usage of the protocol stack will be analysed. The ONE5G[143] project studies RAN KPIs in the context of optimization of the end-to-end 5G performance.

The 5G-MoNArch project is aimed at implementation architecture for 5G, including the issues of underlying network slicing, cross-domain management and run-time optimization. Hence, the project has provided a broad set of KPIs [44] grouped into:

- *General KPIs:* typical, 5G service-oriented requirements;
- *Resilience and Security KPIs:* end-to-end reliability, telco cloud reliability, service restoration time, security threats identification, security failure isolation;
- *Resource Elasticity KPIs:* availability, cost efficiency gain, elasticity orchestration overhead, minimum footprint, multiplexing gain, performance degradation function, rescue and reparability, resource consumption, resource savings, response time, resource utilisation efficiency, service creation time, time for reallocation of a device to another slice, and;

[141] ETSI NFV: *https://www.etsi.org/technologies/nfv* [Accessed April 2019].

[142] 5GENESIS ("5th Generation End-to-End Network, Experimentation, System Integration and Showcasing") project (GA No. 815178). Main website: *https://5genesis.eu/* [Accessed April 2019].

[143] ONE5G ("E2E-aware Optimizations and advancements for the Network Edge of 5G New Radio") project (GA No. 760809). Main website: *https://one5g.eu/* [Accessed April 2019].

- *Application-specific KPIs:* frame rate judder, the maximum number of simultaneously active IoT devices, task success rate, time on task, use of search *vs* navigation, etc.

In the context of use case-oriented KPIs, the 5GCHAMPION[144], 5GCAR[145] and TRIANGLE[146] projects are worth a mention. The 5GCHAMPION has been focused on 5G use cases used during the 2018 Winter Olympics in Korea. The 5GCAR has defined automotive (service) KPIs and also includes communication network (radio) KPIs. The TRIANGLE project has defined and implemented a framework to test and benchmark 5G applications, devices, and services. All the mentioned projects and initiatives follow the ITU-R or 3GPP visions of requirements definitions of the end-to-end communication service.

So far, most approaches to 5G are unaware of the performance and quality of network slicing. There are some, above-mentioned SDOs and research projects efforts related to network slicing, but the issue has not been solved yet. This has motivated us to formulate a small but representative set of KPIs that can be used in order to assess the impact of the implementation of certain networking solution (e.g. EPC) as a network slice. It should be noted that *on the one hand,* the virtual implementation of the network solution brings some benefits but, *on the other one,* it creates some problems. Having in mind that the number of slices can be huge, the number of network-slicing related parameters must be kept to the minimum in order to minimize the overhead related to their collection, calculation, and interpretation. The proposed KPIs for network slicing are described in the next section.

4.3.2 Network slicing KPIs set proposal

The following part focuses only upon these KPIs that are related to network slicing. The KPIs related to the solution, which is implemented as a network slice, are out of the scope of this analysis because they should be exactly the same as defined for the non-sliced implementation of the solution. For example, if 4G EPC is implemented as a slice, KPIs that concern EPC implementation in a non-sliced (or virtualised) environment are applicable. Therefore, the main focus is only on the parameters that are defined by 3GPP as common for different network functions.

To define the network slicing KPIs, it is necessary to use a certain functional model of network slicing as well as its implementation. The 5G-DRIVE project follows the NGMN functional approach [20] with some extensions and the ETSI MANO approach for slice orchestration. KPIs for slice run-time and life-cycle operations are analysed separately. Moreover, a separate analysis for RAN slicing is provided, as the technology virtualisation differs from other domains. The proposed KPIs not only deal with a single slice but also include the impact of the multiple slices on infrastructure and single slice operations. KPIs for multi-domain slices are also analysed. In the case of multiple domains, the approach with local (per domain) orchestrators and stitching of domain slices for obtaining the end-to-end slice is considered. Such an approach requires a master orchestrator on top of local orchestrators and well-known domain-slices. In comparison to the multi-domain orchestration that uses a single slice blueprint, single orchestrator and the capability of allocation of resources of all domains, such orchestration is faster. Moreover, local orchestrators can handle domain-specific issues that are of premium importance (e.g. for RAN).

---

[144] 5GCHAMPION ("5G Communication with a Heterogeneous, Agile Mobile Network in the Pyeongchang Winter Olympic Competition") project (GA No. 723247). Main website: *http://www.5g-champion.eu/* [Accessed April 2019].

[145] 5GCAR ("Fifth Generation Communication Automotive Research and innovation") project (GA No. 761510). Main website: *https://5gcar.eu/* [Accessed April 2019].

[146] TRIANGLE ("5G Applications and Devices Benchmarking") project (GA No. 688712). Main website: *https://www.triangle-project.eu/* [Accessed April 2019].

In this concept, the 5G-DRIVE context partly uses and updates the KPIs defined by 3GPP and performance measurements specified by ETSI. In comparison to the existing approaches, a set of KPIs is provided, which are calculated simpler (only the threshold exceeds are reported) but provide the same sufficient information about the behaviour of the network slicing. In case of detected issues, the management and orchestration system is assumed to "trigger" some action that will include more detailed monitoring and solving the issue. The proposed KPIs deal with some performance-related indicators that typically change dynamically, therefore the 5G-DRIVE does not address slice parameters that are static. The proposed KPIs can be split into slice run-time and slice life-cycle management related.

4.3.2.1 Slice run-time KPIs

Slice run-time KPIs concern performance of a network or a service that is implemented as a slice and typically are identical as in case of non-sliced implementation of the network or solution. The only new mechanisms that are slice-agnostic in the virtualised implementation are related to the usage of virtual resources by a slice and orchestration operations related. One of the key operations on resources is resource scaling according to their usage. The 5G-DRIVE focuses on three types of virtual resources, namely connectivity, computing, and memory. In ETSI NFV framework, memory (i.e. RAM and swap space) and disc measurements are performed separately – thus introducing a single, synthetic parameter related to the usage of all kinds of memory.

Assuming an ETSI MANO compliant orchestrator is used, which has the capability of the virtual resources' dynamic allocation according to slice needs (resource scaling), the proposed KPIs concerning resource usage by a slice are the following:

- **KPI-R1:** Connectivity Resources Underutilisation (ConRu). The KPI is calculated periodically as a number of virtual links of a slice with utilisation under Th_{lo} of link capacity during the observation time T_o. In properly allocated connectivity resources, such a situation should be rare. The abundant connectivity resources have no negative impact on slice behaviour, but they affect the overall efficiency of the resources' usage.
- **KPI-R2**: Connectivity Resources Overutilisation (ConRO). The KPI is calculated periodically as a number of virtual links of a slice which utilisation is over Th_{hi} of link capacity during the observation time T_o. In properly allocated connectivity resources, such situation should be rare. The lack of connectivity resources leads to increased transmission delay and packet loss. Therefore, it degrades slice behaviour.
- **KPI-R3**: Computing Resources Underutilisation (ComRU). The KPI is calculated periodically as virtual CPU utilisation of each VNF of a slice and represents the number of VNFs which computing utilisation is under Th_{lo} during the observation time T_o. The abundant computing resources have no negative impact on slice behaviour, but they affect the overall efficiency of the infrastructure resources usage.
- **KPI-R4**: Computing Resources Overutilisation (ComRO). The KPI is calculated periodically as virtual CPU utilisation of each VNF of a slice and represents the number of VNFs which computing utilisation is over Th_{hi} during the observation time T_o. The lack of computing resources leads to increased processing time. Therefore, it degrades slice behaviour.
- **KPI-R5**: Memory Resources Underutilisation (MemRU). The KPI is calculated periodically as virtual memory utilisation of each VNF of a slice and represents the number of VNFs with memory utilisation of under Th_{lo} during the observation time T_o. The abundant memory resources have no negative impact on slice behaviour, but they affect the overall efficiency of the infrastructure resources usage.
- **KPI-R6**: Memory Resources Overutilisation (MemRO). The KPI is calculated periodically as virtual memory utilisation of each VNF of a slice and represents the number of VNFs with memory utilisation of over Th_{hi} during the observation time T_o. Lack of memory resources leads to increased processing time. Therefore, it degrades slice behaviour.

- **KPI-R7**: Overall Connectivity Resources Underutilisation (OConRU). This KPI is defined as the aggregation of ConRU KPIs of all slices.
- **KPI-R8**: Overall Connectivity Resources Overutilisation (OConRO). This KPI is defined as the aggregation of ConRO KPIs of all slices.
- **KPI-R9**: Overall Computing Resources Underutilisation (OComRU). This KPI is defined as the aggregation of ComRU KPIs of all slices.
- **KPI-R10**: Overall Computing Resources Overutilisation (OComRO). This KPI is defined as the aggregation of ComRO KPIs of all slices.
- **KPI-R11**: Overall Memory Resources Underutilisation (OMemRU). This KPI is defined as the aggregation of MemRU KPIs of all slices.
- **KPI-R12**: Overall Memory Resources Overutilisation (OMemRO). This KPI is defined as the aggregation of MemRO KPIs of all slices.

The increase of above-mentioned KPIs can be observed when the orchestrator part responsible for resources allocation (VNFM, VIM) is overloaded or driven by not optimal resource allocation algorithm.

There exist multiple RAN virtualisation approaches. In most of them, limited RAN virtualisation is used, then other techniques of allocation of resources to slices are often used, for example, a double-level scheduler responsible for Resource Blocks allocation. One level of the scheduler is used for splitting the radio resources (Resource Blocks) between slices; another one is responsible for the allocation of resources to slice users. The intra-slice scheduler KPIs can be linked with classical RAN KPIs. The above-proposed KPIs related to connectivity can be used, whereas the memory- and computing-usage KPIs make sense only in case of virtualised implementation of RAN nodes.

4.3.2.2 Slice lifecycle KPIs

The list of the proposed KPIs includes:

- **KPI-L1**: Slice Deployment Time (SDT) is a parameter that describes the interval between the slice deployment request and the moment in which slice is ready for operation. The problem with this parameter is that this interval depends on “slice template (blueprint) complexity”, the performance of orchestrator, and the allocation of resource time by the virtualised infrastructure. The slice complexity may deal with the footprint size of VNFs, their interconnection topology, amount of configuration parameters. Therefore, in a generic case, it is impossible to define the required value of SDT. It can be noted that SDT can be critical for some network slices, for example in case of on-demand or short-lived ones, but much less critical for long-lived slices.
- **KPI-L2**: Slice Deployment Time Scalability (SDTS) is a measure of scalability of slice deployment operations. To evaluate the scalability, it is possible to send N slice deployment requests of the same slice template and calculate SDTS in the following way:

$$\mathrm{SDTS} = \frac{\mathrm{GSDT}}{\mathrm{N} \times \mathrm{SDT}} \qquad \text{Equation 9}$$

  where GSDT is the overall time for the deployment of N identical slices and SDT is the deployment time of a single slice (as defined above). It is hard to define the N value *a priori*. If the N value is too big, then there can be a problem with the availability of the requested resources. If it is too small, then the obtained result may not express the scalability of the orchestration well. It can be recommended to calculate the SDTS parameter for N = 10. The SDTS is expected to be greater than 1.
- **KPI-L3**: Reconfiguration Execution Time (RET). There are two reasons for run-time slice reconfiguration. The first is driven by NFVO, which decides to move an NFV from one data centre to another. Such reconfiguration is not driven by orchestrator operator but is an autonomic decision of NFVO. The network slice template is unchanged. Such an operation is

executed in the background. Therefore, no KPIs in such case are needed. Another kind of reconfiguration is a reconfiguration that is needed because of changes of the slice blueprint, for example by adding or removing VNFs. Such operation is driven by OSS, and its completion should be reported to OSS. Having in mind the operation, a KPI called Reconfiguration Execution Time (RET) is proposed, which is defined as an interval between slice reconfiguration request and slice reconfiguration completion. It has to be noted that RET is not agnostic and depends on the slice template as well as a number of modifications (number of VNFs, number of links to be reconfigured, number of operations related to VNFs reconfiguration). Hence, the expected value of this KPI cannot be defined, but it should be kept as low as possible. RET has a very specific usage in case of RAN. A RAN can be virtualised partly and may be based on PNFs. The reconfiguration of PNFs that already are in use by a specific slice as a change of allocated resources has been already described. In case of RAN slice reconfiguration, the change in the area that is served by a slice, by adding and/or removing of some RAN nodes to a slice, should be considered. Such operation can still use the RET KPI. However, the value of RET can be different for the different number of RAN nodes involved in the reconfiguration process. It has to be noted, however, that the process of slice coverage modification has not been addressed by 3GPP, yet.

- **KPI-L4**: Slice Termination Time (STT) is a parameter that describes the interval between the slice termination request and the moment in which all slice allocated resources are released. If the time is long, it decreases the efficiency of the infrastructure resources usage.

4.3.2.3 KPIs for multi-domain slicing

The end-to-slice can be created not only in a single domain but also in multiple domains (i.e. administrative, technological or orchestration domains). The creation of the end-to-end slice in multiple administrative domains deals mostly with business-related issues as well as the definition of the operations that are allowed in the non-owned domains. The technological domain in the context of network slicing may imply specific orchestration or management functions, for example by the use of specialised hardware, legacy subsystems or special virtualisation techniques. In the context of mobile networks, such specific operations are typical for RAN and may concern both, nodes of RAN and data transport (especially the front-haul). In general, there are two options for creating end-to-end slices in multiple domains.

The first one lies in the use of single orchestrator that is able to orchestrate the resources of all domains. Another one lies in the usage of per-domain orchestrator and higher-level entity in order to coordinate the behaviour of domain level orchestrators and to provide the end-to-end operations. Such case has the ability to handle domain-specific orchestration issues, but it requires well-defined types of local slices and their descriptors that can be selected for defining the end-to-end slice. From the KPI calculation point of view, the first case (i.e. single orchestrator-based) has no impact on KPI calculation – they are computed in the same way as in the single-domain case described above.

© 2018 - 2021 5G-DRIVE Consortium Parties

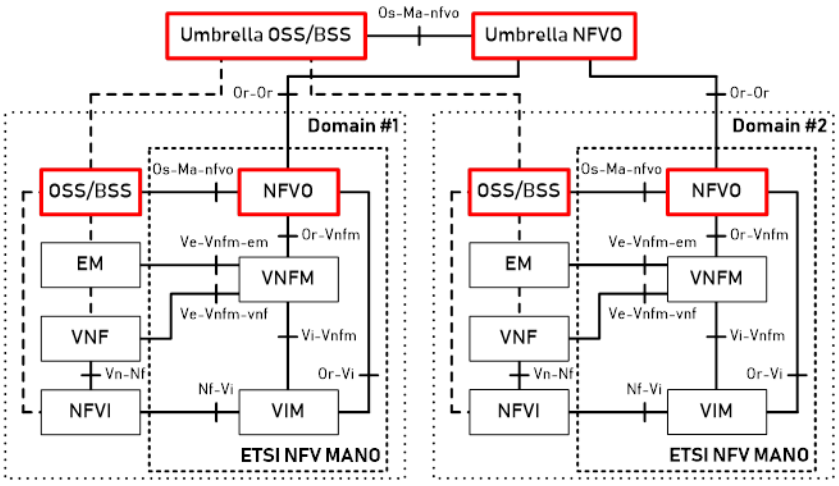


Figure 46: Multi-domain management and orchestration architecture for end-to-end KPI calculation.

The second case is much more complicated. Here, the 5G-DRIVE proposes computation of domain-level KPIs by local OSS/BSS and passing this information to the end-to-end orchestrator (Umbrella NVFO, which may also play a role of a local domain NFVO, cf. [45]). This orchestrator exposes overall KPIs as well as domain-level KPIs to the end-to-end slice operator and slice tenants. The overall KPIs are computed using the operations on domain-level KPIs of all domains. In some cases, the overall KPIs can be obtained by summing of domain-level KPIs in other cases other operations on-local KPIs are applicable. The end-to-end slice setup time is calculated as an interval between the first domain-level slice set-up request and last domain level confirmation of slice deployment. Please note that in opposite to some multi-domain network slicing concepts, the 5G-DRIVE sees a significant role of domain-level OSS/BSS. The concept is illustrated in Figure 46. The Umbrella OSS/BSS can collect and expose not only the network slices' KPIs but also KPIs of "sliced" solutions.

4.3.3 Network slicing KPIs computation in the NFV MANO case

The KPIs described in Section 4.2, in order to be calculated in the MANO environment, require some information about the resource allocation, usage, as well as the occurrence of certain operations and their completion time. In fact, the required for KPIs data may be categorised in the following way:

- Information related to computing, memory, storage and connectivity resources allocated to VNFs and consumed by them;
- information about initiation and completion time of selected NFVO procedures that are driven by OSS/BSS;
- information about VNFM operations (initiation, completion).

In the presented approach, the 5G-DRIVE proposes to use the OSS/BSS of the standard NFV/MANO architecture to calculate and collect the network slicing KPIs. For the collection of information required for KPI calculation, the OSS/BSS has to interact with other components of the MANO architecture. As illustrated in Figure 46, the VIM exposes the information about the underlying NFVI at the reference points Vi-Vnfm and Or-Vi to higher-level MANO entities, VFNM and NFVO. These entities are able to understand, correlate and further enrich this received information in the context of the installed network service description. The OSS/BSS can directly use the Os-Ma-Nfvo reference point of NFVO, for the purpose of Network Service Life-cycle Management (instantiating, scaling, updating, healing, terminating, deleting, etc.); Performance Management (management of performance management jobs and thresholds); Fault Management (management of subscriptions to notifications, querying alarms lists, acknowledging alarms), and; NFVI Capacity Information (querying and notifications about underlying infrastructure capacity and its shortage). Hence, the OSS/BSS is able either to determine the life-cycle operations performance based on a request-response time interval or get directly the subscribed or requested run-time

performance/fault/capacity information. While the information exchange between NFVO and OSS/BSS is at the level of Network Service Instance, the individual VNF's Element Manager (EM) is partially able to exchange similar information with the VNFM at the level of its VNF/VNFCs via the reference point Ve-Vnfm-em and to share further this information with its own OSS. Here, Performance Management and Fault Management interfaces are available. In the case of multi-domain operations, the reference point Or-Or between the producer-NFVO and the Umbrella NFVO copies the definition of Performance, Fault and Life-cycle Management interfaces exposed at Os-Ma-Nfvo. Additionally, the Network Service Instance Usage Notification Interface at Or-Or provides the awareness of the delivered Network Service utilisation status to the producer-NFVO, which can accordingly adapt (e.g. the activities related to performance or the fault reporting for the specific Network Service (NS)).

Specifically, ETSI GS NFV-IFA 027 concerns MANO performance issues and describes mechanisms of premium importance for KPIs calculation:

- VIM uses reference points Vi-Vnfm and Or-Vi to report NFVI-related performance indicators to VNFM and VNFO, *respectively*. The performance metrics include mean/peak usage of virtual CPU, memory, disk, and virtual storage, number of incoming/outgoing bytes/packets on the virtual computer (split per virtual interface) or virtual network (split per virtual port);
- VNFM maps the above-mentioned information from VIM to specific VNFs/VNFCs and exposes the performance measurements at reference points Ve-Vnfm-em (for VNFs/VNFCs) and Or-Vnfm (for VNFs only). These are VNF/VNFC-specific mean/peak usages of virtual CPU, memory, disk and virtual storage, numbers of incoming/outgoing bytes/packets at VNF internal/external connection points;
- The performance measurements produced by NFVO can be transferred to OSS/BSS via the reference point Os-Ma-Nfvo. They include numbers of incoming/outgoing bytes/packets at Network Service border interfaces.

According to ETSI, MANO enables charging of two categories: Usage Events and Management and Orchestration Events. Both types of events can be used in order to calculate KPIs. The Usage Events can be used for resource usage monitoring (computing, storage, networking). The VIM is responsible for such monitoring. The VNF instance monitoring (for example the VNF Instance scaling) is done by NFVO/VNFM, and the Network Service Instance is monitored by NFVO. The mentioned recommendation places the charging-related entities as a part of NFVI, VIM, and OSS/BSS. These functions are focused on averaged usage, but the continuous monitoring of resources is possible as well.

The proposed lifecycle KPIs can be obtained using the interaction between the OSS/BSS and NFVO. Hence, the OSS/BSS able to determine clearly both the beginning and the end of the procedure, also in case of disturbances of intra-MANO communication (e.g. OSS/BSS is notified about the delay of procedure execution due to the need of retrying). There are two possible ways of calculation of these KPIs: (i) based on events logging in the on-board OSS/BSS log – each event is logged with a time-stamp, and correlated search of beginning/finishing event for specific procedure is sufficient; (ii) the API for OSS/BSS-MANO communication will typically use the time-out mechanism and the time-out timer will be implemented – its value at the end of the procedure may be instantly passed to the Performance Management engine of the OSS/BSS. The OSS/BSS operations can be supported by EMs of VNFs in order to increase KPIs calculations scalability.

Table 5: Calculating KPIs using MANO.

KPI	MANO role in KPI calculation
ConRU, ComRU, ConRO ComRO MemRO, MemRU	The KPIs can be calculated using information obtained in several ways: (i) VNFM which produces the measurements on the basis of information from VIM, shares them with NFVO, which can further pass them to OSS/BSS, where KPIs will be calculated, and; (ii) EM can receive the reports for its VNF/VNFC from VNFM and

KPI	MANO role in KPI calculation
	pass them to OSS/BSS for KPIs calculation or calculate KPIs locally and pass the results to OSS/BSS.
OConRU, OConRO, OComRU, OComRO, OMemRO, OMemRU	These KPIs are calculated by OSS/BSS as aggregates of respective KPIs of single slices. The operation concerns all active slices.
SDT, SDTS, RET, STT	The operations that are related to these KPIs are triggered by OSS/BSS and executed by NFVO. The NFVO reports their completion to OSS/BSS. Therefore, computation of these KPIs is based on "request to response" time.
Multi-domain KPIs	The multi-domain OSS/BSS obtains all KPIs from domain-level OSS-es and calculates multi-domain KPIs.

4.4 Dynamic Network Slicing

As noted by the NGMN 5G White Paper[147], the success of 5G technology will be based on the ability of operators to provide multiple solutions for all requirements and also to the ability to provide all stakeholders with a unique solution tailored to their specific needs. Unfortunately, today's network architectures are not up to the challenge. The current "one-size-fits-all" approach to wireless networks for all use cases and services to every device everywhere is no longer viable. It does not offer the level of adaptability needed to meet the performance expectations for new and legacy use cases, services, business models, infrastructure usage approaches, and radio access needs that will emerge with 5G.

Among the critical challenges for future 5G network operators is about allocating network resources by maximizing their advantages. Within this framework, life-cycle management of the network slices becomes a critical problem to be taken into account and solved. Aiming to billet the maximum possible number of diversified service requests, 5G network operators have to deploy virtual network functions and allocate network resources quickly to "structure" related network slices[148]. Moreover, they have to scale slices dynamically, according to the varying service load of the underlying infrastructure(s). In contrast, although a network operator has the maximum control over his network slices, it is essential that one slice may still need to perform some sort of control over itself, so that to improve the service quality. Therefore, the dynamic network slicing technique has to consider how to "open" partial permissions to each slice to configure and manage it without raising security[149] issues. In addition, management of network slice has to take place automatically[150] in order to avoid manual efforts and errors.

[147] Next Generation Mobile Networks (NGMN) Alliance: "5G White Paper". Available at: *https://www.ngmn.org/fileadmin/ngmn/content/images/news/ngmn_news/NGMN_5G_White_Paper_V1_0.pdf*

[148] See: A. Galis and C. Lin (2017, March): Towards 5G Network Slicing - Motivations and Challenges. *IEEE 5G Technology Focus*, vol. 1, no. 1. Available at: *http://5g.ieee.org/tech-focus/march-2017#networkslicing*

[149] Next Generation Mobile Networks (NGMN) Alliance (2016, April): "*5G Security Recommendations Package#2: Network Slicing*". Available at: *https://www.ngmn.org/fileadmin/user_upload/160429_NGMN_5G_Security_Network_Slicing_v1_0.pdf*

[150] Network slicing deployment will benefit from network automation. The automation in rolling network slices will be subject to normal learning curve that new technologies are commonly experiencing. So it can be expected that the degree of automation of deployment of a network slice will further evolve over time.

5G dynamic network slicing[151] can be used to ensure that end-to-end performance meets customer expectations, as well as service and application requirements (Figure 47). To leverage network slicing properly, the individual segments (radio access network, transport, metro, core, edge cloud, central cloud), which were formerly treated separately, must be examined *as a whole*. And, performance optimization must be adapted and coordinated, across the entire network.

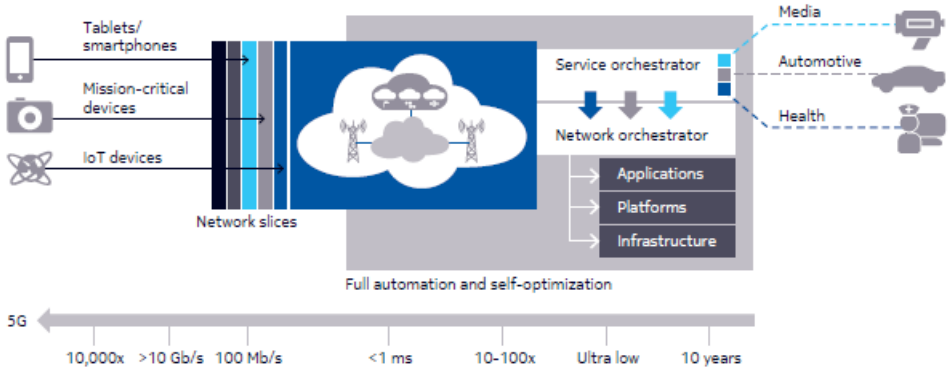


Figure 47: 5G network slicing addressing a variety of industry and multi-tenancy use cases with multiple independent instances on one physical network. [Source: Nokia (2016): "Dynamic End-to-End Network Slicing for 5G- White Paper"].

An effective 5G dynamic network slicing process must[152]: (i) Support a variety of business models, including industry-specific models and, potentially, the multi-tenancy context[153]; (ii) significantly reduce new service creation and activation times; (iii) provide extreme agility in the network to meet diverse service needs; (iv) provide massive elasticity in the network to meet very dynamic traffic demands; (v) exploit analytics and context to adapt services and networks predictively and in real time; (vi) enable an open services ecosystem where different parties can cooperate to introduce innovative services tailored to specific user or industry demands, (vii) expose actionable network insights to application and content providers, enterprises, and industry verticals; (viii) provide full programmability to enable easy integration of new network capabilities, extension of existing capabilities and easy creation of new services and business models; (ix) intelligently manage and orchestrate resources and capabilities for dynamic (re-)configuration of the network to meet end-to-end performance, and; (x) support a high level of automation powered by advances in analytics, and machine learning.

The aim proposed by the context of dynamic network slicing is to further extend the capabilities of 5G networks so that for the latter to be able to effectively "address" a variety of related use case and possible applications[154]. This context permits market telecom operators to design, deploy, customize, and optimize the various network slices that are running on the common network infrastructure. In

[151] Further related information also see, *inter-alia*: International Telecommunication Union - Telecommunications Standardization Sector (ITU-T) (2008, December): ITU-T Recommendation Y.3112 (12/2008): "Framework for the Support of Multiple Network Slicing". Available at: *https://www.itu.int/rec/T-REC-Y.3112-201812-I/en*

[152] See: Nokia Oyz (2016): "Dynamic End-to-End Network Slicing for 5G- White Paper", Espoo, Finland. Available at: *http://www.hit.bme.hu/~jakab/edu/litr/5G/NOKIA_dynamic_network_slicing_WP.pdf*

[153] See, *inter-alia*: K. Samdanis, X. Costa-Pérez, and V. Sciancalepore (2016, July): From network sharing to multi-tenancy: The 5G network slice broker. *IEEE Communications Magazine*, vol. 54, no. 7 pp. 32-39.

[154] See, *among others*: European Telecommunications Standards Institute (ETSI): ETSI GR NGP 011 v1.1.1 (2018-09): "E2E Network Slicing Reference Framework and Information Model".

© 2018 - 2021 5G-DRIVE Consortium Parties

particular, beyond its dependence upon SDN and NFV[155], the dynamic network slicing similarly depends on end-to-end orchestration and analytics.

5G network slicing[156] brings flexibility while increasing the complexity of network management. Thus, there is a need to design automated management schemes to address such complexity. The lifecycle management is responsible for slice creation, reconfiguration, deletion, etc. It involves orchestrating and allocating infrastructure resources. Therefore, it is crucial to improve resource utilisation of 5G infrastructure. Vendors usually consider the term dynamic network slicing because operators will be able to dole out these network slices quickly and on demand. With this type of network slicing, operators can rapidly start deployments, architecture types, and performance thresholds for distinct use cases or service groups.

The dynamic network slicing concept practically extends the slicing vision of the NGMN[157] and adequately covers all requirements, already identified above. In fact, it partitions a common network infrastructure into multiple, logical, end-to-end, virtual network instances or slices with several key characteristics listed as follows:

- The slices support a group of services, use-cases and related business models with related requirements. For example, an operator can run enhanced broadband slices to offer a variety of broadband services to its customers, that include web browsing, audio and video streaming, and chat.
- The slices are built with only relevant network capabilities that "match" the needs of the corresponding supported service, use case or business case. For example, an ultra-low latency capability can be created for a slice supporting ultra-low latency use cases. The capabilities in the slice are not restricted to the user plane. The slices can also control and manage plane-relevant capabilities, such as a dynamic video stream controller or a specific type of billing application relevant to the business case.
- The slices are dynamic in runtime. They comprise an automation framework that uses real-time analytics and monitoring for efficient use of network and cloud resources, and optimization for the dynamic needs of services or dynamic traffic demands.

To enable an effective approach, the dynamic network slicing leverages key technology advancements in: (i) Distributed cloud infrastructure and cloud native applications; (ii) NFV; (iii) E2E orchestration; (iv) SDN and programmable networking; (v) network big data, analytics, and machine learning; (vi) services oriented architectures, and; (vii) intent based network programming.

4.4.1 Dynamic Network slicing as Enabler for End-to-End Performance and for a fully Programmable Network

Subscribers measure service satisfaction based on end-end performance, they are not aware or concerned if an application is running well in the radio, transport, core or the cloud.

To meet future performance expectations, the individual network segments (radio access network, transport, metro, core, edge cloud, central cloud), which were formerly treated completely

[155] For this specific case also see, *inter-alia*: M.-P. Odini (2017, September): SDN and NFV Evolution Towards 5G. *IEEE Softwarization Magazine*. Available at: *https://sdn.ieee.org/newsletter/september-2017/sdn-and-nfv-evolution-towards-5g*

[156] For example, also see the context proposed in: 3GPP TS 28.531: "Provisioning of network slicing for 5G networks and services: Detailed specification of network slice provisioning/Network slice management".

[157] See, *for example*: S. Costanzo, I. Fazzari, N. Aitsaadi and R. Langar (2018): *Dynamic Network Slicing for 5G IoT and eMBB services: A New Design with Prototype and Implementation Results*. In IEEE (Ed.), Proceedings of the 3rd Cloudification of the Internet of Things (CIoT 2018). Paris, France, July 02-04, 2018.

separately, must be re-evaluated. Their performance must be adapted and coordinated to deliver a specific flow, for a specific user, at specific time.

Dynamic network slicing offers an effective way to do that, meet all use case requirements, and exploit the benefits of a common network infrastructure. It enables operators to establish different deployments, architectural flavors, and performance levels for each use case or service group and run all network implementations in parallel.

With dynamic network slicing, operators could create a fully programmable network architecture which suits the requirements of various use cases, subscriber types and apps. They could allocate dedicated and optimized end-to-end virtual network functions and physical resources for each use case or group. And within a slice, the functional elements could be instantiated according to the needs of a specific service offering. This is achieved by analyzing service requirements and identifying KPIs. With this information, the right network resources could be allocated at the right time and in the right place based on performance requirements, the QoE needed, and the relevant service quality indicators that must be monitored and assured. For example, latency-critical services could be allocated functions in the radio access or edge cloud to minimize delay. But to improve efficiency, these same functions could also be allocated in the central cloud for other services.

4.4.2 NFV and SDN as Enablers for Network Slicing

The dynamic network slicing concept leverages NFV and SDN to create many dedicated end-to-end virtual networks[158]. All end-to-end network slices are created and operated over a common physical infrastructure. A slice is self-contained. It has all the functions and capabilities, appropriately chained together to best meet all the needs of the corresponding services and use-cases. Dynamic network slicing could also be used to support a variety of business cases. For example, “multi-tenancy” in which mobile network tenants could share the same network infrastructure with vertical industry[159] tenants. Each tenant could operate and manage their corresponding network slice by exploiting built-in network slice customization and optimization capabilities. With this approach, the network could be sliced:

- Per type of customer services when an operator wants to optimize dedicated services for end users (e.g. tactile internet or ultra HD video) or wants to address industry-specific customer needs (e.g. e-health, sensor network, high speed train).
- Per tenant when an operator wants to share network costs (e.g.: Network Platform as-a-Service (NPaaS) and Network Infrastructure as-a-Service (NIaaS)).

4.4.3 Management and Orchestrating of Dynamic End-to-End Network Slices

An important component of network slicing is a MANO (MANagement and Orchestration) system that enables full lifecycle management of network slices and associated network resources.

By nature, Network Functions Virtualisation (NFV) changes the way networks are managed. From initial set up to day-to-day operations, NFV management and network orchestration (MANO) fills the

[158] L.M. Contreras, P. Doolan, H. Lønsethagen, and D.R. López (2015): Operation, organization and business challenges for network providers in the context of SDN and NFV. *Elsevier Computer Networks*, vol. 92, pp. 211-217.

[159] Traditionally, vertical industries run their services in a shared network environment upon which infrastructure owners and service providers offer standalone network capabilities including connections, storage, etc. The (dynamic) network slicing paradigm enables supporting the requirements of a network slicing tenant to be met individually. Hence it is anticipated that this type of new business model where network slice instances are leased to industry verticals as a service (i.e.: Network Slicing as-a-Service (NSaaS)) may become a broad norm in the near future.

management role. The NFV MANO is a framework developed by a working group of the same name within the European Telecommunications Standards Institute Industry Specification Group for NFV (ETSI ISG NFV)[160]. Over time, the framework became more commonly referred to as just NFV management and orchestration. NFV MANO is broken up into three functional blocks:

- *NFV Orchestrator (NFVO)*: It is Responsible for on-boarding of new network services (NS) and virtual network function (VNF) packages; NS lifecycle management; global resource management; validation and authorization of network functions virtualisation infrastructure (NFVI) resource requests.
- *VNF Manager (VNFM)*: This block oversees lifecycle management of VNF instances; fills the coordination and adaptation role for configuration and event reporting between NFVI and Element/Network Management Systems (EMS/NMS).
- *Virtualised Infrastructure Manager (VIM)*: This block controls and manages the NFVI compute, storage, and network resources.

The slicing management and orchestration functional component is a key part of the network slicing framework. It manages and orchestrates the physical resources, virtual resources and network slices in a way the NFV-MANO will be built in. Slicing MANO can have sub-functions such as infrastructure manager, VNF manager, and slices orchestrator. The tasks of slicing MANO include: (i) creating and managing Virtual Machine (VM) instances by using the infrastructure resources; (ii) mapping network functions to virtual resources and connecting network functions to create service chains; (iii) managing the life cycle of network slices by interacting with the application and service layer, i.e. automated creation of service-oriented slices and dynamic maintenance by monitoring service requirements and virtual resources. In fact, the management of network slices may be performed at two levels, which are inter-slice management and intra-slice management. Inter-slice management is performed by the slice orchestrator while intra-slice management is implemented by running a virtual manager function as a part of the slice.

The ETSI NFV[161] standard defines lifecycle management of “Network Services” (“NSs”) that can be reused in the lifecycle management of network slicing[162]. It is possible to derive and find similarities[163] between 3GPP network slicing concept and the ETSI NFV concepts of ETSI NFV NS. The latter is composed of VNFs and PNFs, along with the virtual links or interconnections between them.

An E2E network slice orchestrator is envisioned to manage all aspects of network slicing. Such an orchestrator will help to manage the entire life cycle management of a slice, from design and creation to operation and optimization.

Automation for network slices is a key strategy to realize operational goals. It starts with transferring the software for network slices, which may come from vendors, to the operator, followed by initial testing of the slices in the operator’s network[164]. Various suppliers (i.e. software developers)

---

[160] It is the ETSI-defined framework for the management and orchestration of all resources in the cloud data center. This includes computing, networking, storage, and virtual machine (VM) resources. The main focus of NFV MANO is to allow flexible on-boarding and sidestep the chaos that can be associated with rapid spin up of network components. Also see: *https://www.sdxcentral.com/networking/nfv/definitions/who-is-etsi-network-functions-virtualization-nfv/*

[161] The meaning of the term “Network Service” in ETSI NFV is different from the meaning of this term in 3GPP; with the latter usually understood as the service supplied to specific user.

[162] For example, also see the context proposed in: 3GPP TR 28.801 V15.1.0 (2018-01): “Telecommunication management; Study on management and orchestration of network slicing for next generation network – Release 15”.

[163] See, *for example*, the case of the H2020/5G-PPP (Phase-1) SESAME project (under Grant Agreement No. 671596) at: *http://www.sesame-h2020-5g-ppp. eu/*. Also see the case of the H2020/5G-PPP (Phase-2) 5G ESSENCE project (under Grant Agreement No. 761592) at: *http://www.5g-essence-h2020.eu/*

[164] Also see the discussion and the dedicated context as presented in: L.M. Contreras, and D.R. López, (2018): A Network Service Provider Perspective on Network Slicing. *IEEE Softwarization Magazine: A Collection of Short Technical Articles.* Available at:

electronically deliver updates to different components that make up a network slice to the operator, where they are automatically integrated with any operator updates or configuration changes. Automation then continues with deploying the network slices in the production network and further controlling the scaling of slices and performing longer term maintenance on the slices.

Slice management and orchestration can be applied at different levels (i.e. both across slices and within the slice) and at different times. Correspondingly, network resources need to be allocated to a slice based on a trade-off between guaranteeing resources to an individual slice and the advantage of pooling resources from all slices. Finally, automated processes would be used to manage the entire life cycle of all slices, implicating for: (i) Monitoring of network functions, services and loads on virtual and physical nodes with well identified and standardized KPIs; (ii) proactive slicing of policies to support decision matching to customer/tenant expectations based on network analytics, and; (iii) configuration and reconfiguration using inputs from analytics and from monitoring specific performance requirements established by policies.

4.4.4 Dynamic End-to-End Network Slicing for optimizing the entire 5G Network

Dynamic end-to-end network slicing provides an optimal approach to network slicing in 5G networks[165]. It actually "partitions" the same network infrastructure into multiple logical end-to-end network slices. With these slices, operators could support the very diverse and extreme requirements for latency, throughput, and availability needed to deliver 5G services to a wide variety of users, machines, industries, and other organizations[166].

The dynamic network slicing concept will be mainly beneficial in the 5G framework, by allowing dynamic end-to-end network partitioning at all network levels (i.e.: from the RAN to the transport network and to the core network). This will offer to the 5G operators the possibility to promote appropriate deployments and architectures for each business model, thus enabling new use cases.

With this innovative approach, operators can produce different deployments and architectural contexts for each use case or dedicated service group. In addition, operators can "run existing network implementations in parallel, simultaneously on a common network infrastructure. This also promotes new communications possibilities and use cases. Monetizing QoE is another area where operators can benefit from dynamic network slicing. In 5G, it is possible to define QoS based on each application's unique requirements. For example, an operator may create different slices for different Mobile Virtual Network Operators (MVNOs), and each slice may have different QoE values, thereby monetizing the experience.

Following to the above and, in order to "address" a variety of diverse requirements, modern 5G network architectures have to "shift" from the current network of entities to a sort of network of capabilities. In this scope, network models can "shift" from the current network for connectivity to a network for services model. Dynamic network slicing offers an effective way to enable this sort of "shift" and partition a single common 5G infrastructure into multiple logical end-to-end networks. It provides the service agility needed to address diverse: (i) Users (both people and machines or other equipment); (ii) use cases; (iii) requirements for latency throughput and availability. With dynamic network slicing, operators can establish different deployments, architectural flavors, and

---

https://e-archivo.uc3m.es/bitstream/handle/10016/26927/network_IEEES_2018.pdf?sequence=2&isAllowed=y

[165] K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini, and T. Braun (2017, August): Network Slices toward 5G Communications: Slicing the LTE Network. *IEEE Communications Magazine*, vol. 55, no. 8, pp. 146-154.

[166] Also see, *inter-alia*: P. Rost, C. Mannweiler, D.S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker (2017, May): Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks. *IEEE Communications Magazine*, vol. 55, no. 5, pp. 72-79.

performance levels for each use case or service group. And they can run all network implementations simultaneously in parallel.

This dynamic network slicing approach is guided by several key principles, as outlined below.

- *A distributed cloud and virtualisation infrastructure*: The infrastructure to support 5G network and network slices is composed by using confirmed virtualisation technologies, and cloud native applications. Wherever possible, the 5G network functions and capabilities are developed upon a distributed cloud and virtualisation infrastructure. Furthermore, dedicated and purpose-built network entities are only used when necessary.
- *A network of capabilities instead of network of entities:* The 5G network slices are composed of modular network capabilities. Network Capability Units (NCUs) are the abstractions of these network capabilities. The NCUs can have varying degrees of granularity, but they are expected to be modular for easy plug-and-play deployment and operation. The NCUs can potentially be implemented as containerized applications, as native micro-services, or as complete virtualised network functions.
- *A network for services instead of a network for connectivity:* The 5G network slices are designed from the ground up to offer and support classes of services. This is a paradigm "shift" from past networks, which were primarily built to offer connectivity. With this approach, the 5G network slices can have unique capabilities that are required for the supported group of services. Moreover, the capabilities of each slice can be dynamically optimized to satisfy the specific needs of individual services.
- *Easy design and dynamic creation of end-to-end slices:* Dynamic 5G network slices can be easily designed by packaging necessary NCUs into a forwarding graph. Network slices can be created dynamically by orchestrating these forwarding graphs on a distributed cloud infrastructure using proven orchestration and management technologies. A template-based approach can be used to ease the creation of slices and to eliminate routine errors. The templates can be used to create various slices and then each slice can be further customized based on the needs of supported services, use cases, or business models.
- *Dynamic programmability and control:* The 5G network slices support dynamic programmability and control by leveraging SDN principles. And the dynamic programming of network slices can be accomplished either by custom programs or within an automation framework driven by analytics and Machine Learning (ML).
- *Automation of the network operation and optimization:* Creating more network slices inherently adds complexity to network operations and ongoing optimization efforts. Therefore, the dynamic network slices are designed from the ground up to enable maximum possible automation of operations and optimizations. Automation is enabled by analytics, machine learning, network big data and network programmability.
- *End-to-end perspective and approach:* The 5G network slices are composed of capabilities from multiple network segments that span the network from access to core, as well as network applications. This end-to-end perspective is needed to meet the needs of diverse services, use cases, and business models. It eliminates the need to make assumptions about what specific sub-segments in the network will or will not support.

4.4.5 Potential Future Trends in the scope of Dynamic Network Slicing

In order to implement and use network slice functions and operations, there is a need to look at the complete life-cycle management characteristics of network slicing solutions based on the following architectural tenets:

- *Underlay tenet*: This implicates for support for an IP-based underlay data plane the transport network uses to carry that underlay.
- *Governance tenet*: It is a logically centralised authority for all the network slices in a domain.

- *Separation tenet*: This implicates that slices may be independent of each other and may have an appropriate degree of isolation from each other.
- *Capability exposure tenet*: This allows each slice to present information regarding services provided by the slice (e.g. connectivity information, mobility[167], automaticity, etc.) to third parties, via dedicated interfaces and/or APIs, within the limits set by the operator.

In pursuit of solutions for the above tenets with the relevant characteristics within the context of 5G networking, there is a need to address the following challenges and outcomes:

- A *Uniform Reference Model* for Network Slicing that describes all of the functional elements and instances of a network slice. It also describes shared non-sliced network parts.
- *Slice Templates*: Providing the design of slices to different scenarios. This outlines an appropriate slice template definition that may include capability exposure of managed partitions of network resources (i.e. connectivity compute and storage resources), physical and/or virtual network and service functions that can act as an independent connectivity network and/or as a network cloud.
- *Network Slice capabilities*, which are expected to be:
  - Four-dimensional efficient slice creation with guarantees for isolation in each of the Data/Control/Management/Service planes. Having enablers for safe, secure and efficient multi-tenancy in slices.
  - Methods to enable diverse requirements for network slicing including guarantees for the end-to-end QoS of a service within a slice.
  - Efficiency in slicing, specifying policies and methods to realize diverse requirements without re-engineering the infrastructure.
  - Recursion, namely methods for network slicing segmentation allowing a slicing hierarchy with parent–child relationships.
  - Customized security mechanisms per slice.
  - Methods and policies to manage the trade-offs between flexibility and efficiency in slicing.
  - Optimisation, namely methods for automatic selection of network resources for network slicing; global resource views; global energy views; network slice deployment based on global resource and energy efficiency; slice mapping algorithms.
  - Monitoring the status and behaviour of network slicing in a single and/or muti-domain environment; monitoring of network slicing interconnection.
  - Capability exposure for network slicing (allowing openness); with APIs for slice specification and interaction.
  - Programmability and control of network slices.
- *Network slice operations*, which are expected to be:
  - Slice lifecycle management including creation, activation/deactivation, protection, elasticity, extensibility, safety, sizing and scalability of the slicing model per network and per network cloud for slices in access, core and transport networks; for slices in data centres, and for slices in edge clouds.

[167] See, *for example*: F. Zarrar Yousaf, M. Gramaglia, V. Friderikos, B. Gajic et *al.* (2017): "Network slicing with flexible mobility and QoS/QoE support for 5G Networks". In IEEE (Ed.), Proceedings of the 2017 IEEE International Conference on Communications Workshops (ICC Workshops 2017). Paris, France, May 21-25, 2017.

- Autonomic slice management and operation, namely self-configuration, self-composition, self-monitoring, self-optimisation, self-elasticity for slices that will be supported as part of the slice protocols.
- Slice stitching/composition by having enablers and methods for efficient stitching/composition/decomposition of slices: vertically (through service and management and control planes); horizontally (between different domains as part of access, core, edge segments); or a combination of both vertically and horizontally.
- End-to-end network segments and network clouds orchestration of slices.
- Service Mapping, by having dynamic and automatic mapping of services to network slices.

- *Efficient enablers and methods for integration* of the above capabilities and operations.

4.4.6 Potential "Go-to-Market" Strategy and Regulatory Aspects

It is expected that dynamic network slicing will support market growth and related investments[168]. A probable Go-to-Market strategy exercised by the corresponding "market actors" may implicate for three distinct stages. These stages deliver value across the value chain[169], and may happen in parallel depending on local market conditions. These are briefly discussed as follows:

- *Deploy network slicing for internal use*: This will prove the validity of network slicing by using the solution to serve internal customers within an operator or the operator's sister companies. This option offers a low risk opportunity to experiment and to validate the proposition in order to refine it ahead of rollout to commercial customers.
- *Upsell network slicing capabilities to existing enterprise customers*: This will prove value to existing enterprise customers and based on the typical buying behaviour of business customers, upselling network slicing capabilities to these customers ought to be an easier opportunity than targeting new customers. These customers can then become proof points and advocates for the new capabilities.
- *Sell to new enterprise customers*: When commercially ready, slicing will be made more broadly available to enterprise customers who often require a proven solution and seek market validation before they buy.

Network slicing will be a feature of 5G, based upon software defined network and network function virtualisation. In reviewing network slicing, regulation should be considered and views on several "key areas" are provided below. However, from a regulator's perspective they will want to understand what, if anything changes from developments in technology, and how this impacts customers. These changes might impact and/or be constrained by the related regulatory aspects, listed below. At this stage there do not seem to be new regulatory challenges as a result of network slicing. However, the challenge(s) may be on how best to meet existing regulatory obligations as result of network slicing.

- *Net neutrality*: While there is no single definition of "net neutrality", it is often used to refer to issues concerning the optimisation of traffic over networks. Mobile network operators "face" unique operational and technical challenges in providing fast, reliable internet access to their customers, due to the shared use of network resources and the limited availability of

[168] Global System for Mobile Communications Alliance (GSMA) (2017, September): "Smart 5G Networks: Enabled by Network Slicing and Tailored to Customers' Needs". Available at: *https://www.gsma.com/futurenetworks/wp-content/uploads/2017/09/5G-Network-Slicing-Report.pdf*

[169] Also see, *among others*: B. Han, S. Tayade, and H.D. Schotten (2017): Modeling Profit of Sliced 5G Networks for Advanced Network Resource Management and Slice Implementation". In IEEE (Ed.), Proceedings of the 2017 IEEE Symposium on Computers and Communications (ISCC-2017). Heraklion, Greece, July 03-06, 2017.

spectrum. Network slicing may provide options of how mobile operator networks can better meet customer needs.

- *Quality of Service*: The quality of a mobile data service is characterised by a number of important parameters, notably speed, packet loss, delay and jitter. It is affected by factors such as mobile signal strength, network load and user device and application design.
- *Cross-border data transfers*: The global digital economy depends on cross-border data transfers to deliver crucial social and economic benefits to individuals, businesses and governments. When data is allowed to flow freely across national borders it enables organisations to operate, to innovate and to access solutions and support anywhere in the world. Enabling cross-border data transfers can help organisations adopt data-driven digital transformation strategies that ultimately benefit individuals and society. Policies that inhibit the free flow of data through unjustified restrictions or local data storage requirements can have an adverse impact on consumers, businesses and the economy in general. Cross-border transfers of personal data are currently regulated by a number of international, regional and national instruments and laws intended to protect individuals' privacy, the local economy or national security.
- *Illegal content*: Today, mobile networks not only offer traditional voice and messaging services but also provide access to virtually all forms of digital content via the internet. In this respect, mobile network operators offer the same service as any other Internet Service Provider (ISP). This means mobile networks are inevitably used, by some, to access illegal content, ranging from pirated material that infringes intellectual property rights (IPR) to racist content or child sexual abuse material (child pornography). Laws regarding illegal content vary considerably. Communications Service Providers (CSP), including mobile network operators and ISPs, are not usually liable for illegal content on their networks and services, provided they are not aware of its presence and follow certain rules (for example, "notice and take-down" processes to remove or disable access to the illegal content as soon as they are notified of its existence by the appropriate legal authority).

4.5 Multi-provider infrastructure extensions for MANO

The deployment of network slices in a single domain is not realistic and does not exploit the full potential of the technology. Even if some MANO Operators try to provide the services at all layers (E2E slice orchestration, network sub-slice orchestration, infrastructure management and infrastructure operation), the common market of interconnected infrastructure will consist of multiple resources' owners, who are just interested in exposing and selling their resources without any further responsibility. The aggregation and brokering of resources are so needed. The resources of the specific owner may be also exposed to different Infrastructure Brokers or "integrators", i.e. Slice Brokers, who offer Network Services (in ETSI NFV meaning) implemented in at different prices and rules. Due to such approach, existing resources of multiple infrastructure providers can be aggregated; the range can be extended and, in case of the "overlapping" of resources, the resource selection can be based on the price/performance ratio. The slice deployment brokering mechanism lies in the use of the broker model for slice template deployment in a multi Slice Provider environment. Each of the Slice Providers can evaluate which part (if not all) of slice template it can implement and what will be the price of such implementation. As a result of several iterations, such an approach leads to the lowest implementation price of slice template. The 5G-DRIVE context also proposes to add the Service Broker, which serves as single point-of-contact to the customers and it is used for the brokerage of services.

In case of the existence of multiple slice and infrastructure providers, the business model is different than in the previous case and has several brokers. The roles of most of the players are almost the same. However, there are some changes related to the way, in which they interact with brokers. The roles and interactions of the new entities are the following:

The Infrastructure Broker is responsible for the selection of infrastructure resources that are based on Slice Provider's criteria (i.e. price, QoS, area). The Slice Broker provides aggregation of infrastructure resources of different infrastructure providers, if necessary. It exposes the infrastructure in a similar way as in the case of a single domain. However, it provides mechanisms for resource negotiation. In the negotiation process, there can be involved a slice requester (i.e. Slice Tenant, Slice Customer or Service Provider), Slice Provider and the Infrastructure Broker. In the result of the negotiations, there can be selected not only a proper subset of resources but also a slice template. The Infrastructure Broker supports multi-tenancy (i.e. is able to serve multiple Slice Providers). Moreover, it should have mechanisms for infrastructure resource discovery.

The Slice Broker is used by slice requesters (Slice Tenants, Slice Customers or Service Providers) for creation, discovery or termination of a slice (if such operation is allowed). It selects a proper slice template that has to be deployed, as well as deployment rules, and later on, negotiates the deployment of the slice with Slice Providers. The providers may implement a whole slice template or only its part (such limitations can be imposed by limited resources or limited geographic span of the Slice Provider). The process of template split between Slice Providers is therefore iterative and ends when the whole slice template can be deployed or when due to some constraints it is impossible. The Slice Broker performs the role of the higher-level Slice Provider hiding the fact that the slice is deployed using multiple underlying Slice Providers.

The Service Broker is used by Slice Customers for selection or triggering new services. It provides the customers a portal with information about running and available services, as well as their prices. It interacts with Service Providers to provide customer access to services. From the customer point of view, it is a single attachment point to services. However, the Service Broker may redirect the customers to the Service Provider, e.g. for the sake of scalability or business confidentiality.

The presented concept supports multiple Slice Providers. Moreover, the providers can be added dynamically, if the solution allows for the dynamic creation of slice orchestrators. In such a case, the entity that requests the creation of such a slice (Slice Customer, Service Provider or Slice Tenant) will also perform the role of Slice Provider. Such a case, however, needs some advanced technical solutions to be implemented first.

Almost all of network slicing approaches use the ETSI NFV MANO framework as a starting point for the implementation of the network slicing architecture. It has to be noted, *however*, that the MANO framework has not been designed for the multi-provider and multi-tenant environment. Moreover, it has no accounting scheme implemented yet, the OSS/BSS role is not clearly defined, the interaction with Network Service (slice template) or VNF providers is internal to NFVO and not compatible with the DevOps approach. Lack of business-oriented interfaces is the main drawback of MANO in the context of network slicing. The OSS/BSS can be partially exploited for providing multi-tenancy of the orchestrator another possibility is to sue an orchestrator that goes beyond MANO and already supports multi-tenancy, (e.g. some of the functional components of network slicing can be seen as add-ons to MANO).

The ETSI NFV MANO architecture has to undergo many changes to support the multi-domain business concept as described in the previous section. First of all, the interaction with the Infrastructure Broker and appropriate negotiations have to be supported. The broker should be a part of the overall architecture, and its functions should be standardized. Also, NFVO should be modified to interact with Slice Broker. The most important modification of NFVO lies in the capability to evaluate the feasibility and cost of implementation of a part of the slice template only. Such information is needed by the Slice Broker for the lowest-cost implementation of the multi-domain slice.

5 Novel services and cybersecurity

Network function virtualisation (NFV) is one of the cornerstone technologies used within 5G. The European Telecommunication Standards Institute (ETSI) serves as one of its main standardisation drivers. ETSI compliance is a critical component in the 5G-DRIVE framework. The ETSI NFV architecture is used as the starting point for the 5G-DRIVE architecture, aiming to "place" the project in a position where it can contribute to these standardisation activities and align itself to the *de-facto* industry standard. The following subsections discuss the main considerations when designing VNFs for the 5G-DRIVE, focusing on ETSI compliance and optimisations necessary to ensure high performance and high availability.

5.1 ETSI NFV reference architecture

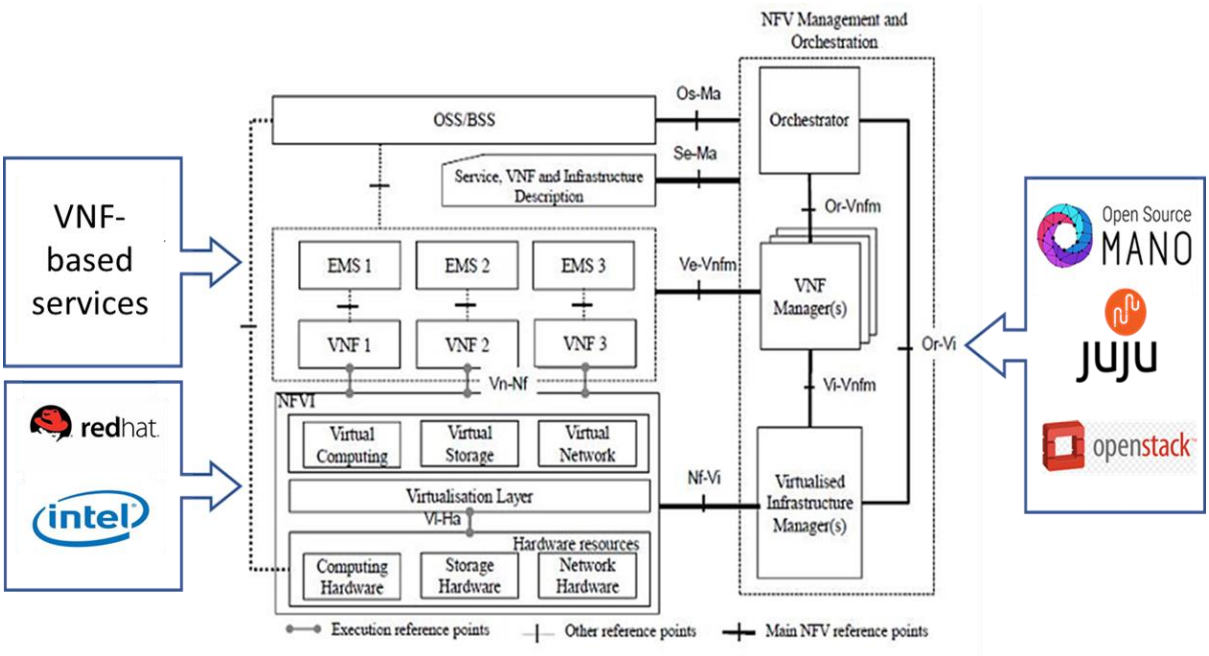


Figure 48: ETSI NFV Standard Architecture and basic technology mapping.

Figure 48 illustrates the standard architecture proposed in ETSI NFV [21] as well as some common technologies mapped to each domain that are currently in consideration for WP5 deployments. It defines three key domains, namely:

- **The Virtual Network Functions (VNFs)** domain explores the aspects related to the instantiated Virtual Network Functions. High availability, elasticity and reliability are key requirements.
- **The Network Function Virtualisation Infrastructure (NFVI)** domain provides the hardware and software infrastructure whereupon the VNFs are executed. It is defined by the back-end software that is necessary to share computing, storage and network resources to the VNFs, the VNF hypervisor, the hardware used, etc.
- **The NFV Management and Orchestration (MANO):** includes all the functions necessary to manage resources, as well as the handling of placement, configuration and lifecycle management of the network services.

5.2 The VNF domain

This section focuses on the implementation of ETSI-compliant VNFs, starting from a description of the general architecture, the VNF interfaces, etc.

5.2.1 VNF & NS composition

Each VNF can be composed by one or more VNF Components (VNFCs) that are interconnected through Virtual Network Links (VLs). The 5G-DRIVE services may consist of one or more VNFs. These NSs will be dynamically deployed to provide specific services. The VNF Orchestrator (NFVO) will be responsible for the orchestration of the VNFs into services and the deployment, management and configuration of the resulting end-to-end network services. An example of a network service (NS) that consists of three different VNFs (VNF1, VNF2 and VNF3) connected through virtual links is shown in Figure 49. As depicted, VNF2 is composed of three VNFCs connected through virtual links that are internal to the VNF.

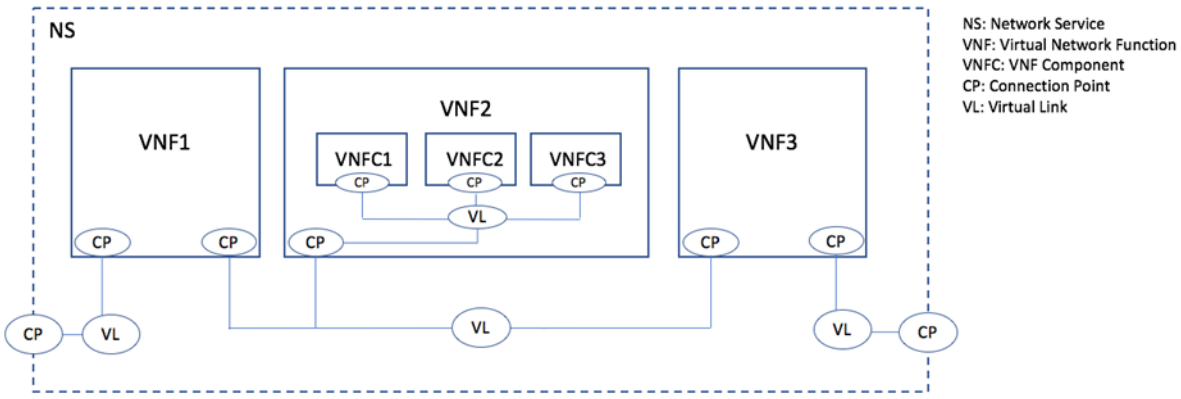


Figure 49: Example architecture of a Network Service with multiple VNFs.

5.2.2 VNF Interfaces

According to ETSI NFV specifications [21], there are five types of interfaces identified relevant to a VNF. As illustrated in Figure 50:

- *SWA-1 interface:* This interface enables communication between various network functions within the same or different network services. The SWA-1 interface can be established between two VNFs, a VNF and a Physical Network Function (PNF), or between a VNF and an End Point. A VNF may support more than one SWA-1 interface.
- *SWA-2 interface:* This interface refers to VNF internal interfaces, for the communication between the different VNFCs of a VNF (i.e. for VNFC to VNFC communication). The type of information exchanged through this interface depends on the function of the VNF.
- *SWA-3 interface:* This interface interconnects the VNF with the NFV management and orchestration layer specifically with the VNF Manager (VNFM). Through this interface, the lifecycle management of the VNF is performed (e.g. instantiation, termination, scaling, etc.). The SWA-3 interface corresponds to the Ve-Vnfm reference point.
- *SWA-4 interface*: This interface is used by the Elemental Management (EM) to communicate with a VNF. It is a management interface used for the runtime management of the VNF to perform functions related to Fulfilment, Assurance and Billing (FAB) as well as Fault, Configuration, Accounting, Performance and Security (FCAPS).
- *SWA-5 interface*: The SWA-5 interface links the VNF with the NFVI and corresponds to the Vn-Nf reference point (as seen in Figure 48). This interface provides access to a virtualised

slice of the NFVI resources allocated to the VNF, i.e. to all the virtual compute, storage and network resources allocated to the VNF depending on the VNF type and its special requirements for resources.

In order to be compliant with the ETSI MANO [34] specifications, the 5G-DRIVE VNFs will support these interfaces.

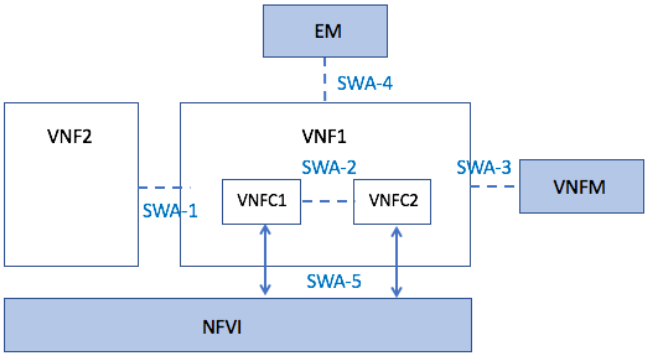


Figure 50: Types of VNF interfaces.

5.3 The NFVI domain

The NFVI domain consists of the hardware, software and networking infrastructure upon which the services are instantiated. This includes host servers, switches and routers, virtualisation software, back-end drivers, etc.

5.3.1 ORION local NFVI

The NFVI Point-of-Presence (PoP) in ORION's Athens site (Figure 51) runs the OpenStack [46] Ocata distribution[170] based on Centos 7.4.1708[171]. The OpenStack controller and a compute node are situated on a single server, thus denoting this an "all-in-one" deployment. The PoP provides networking to the VNFs through OpenStack's Neutron[172] service. All the networking is therefore handled automatically by OpenStack, provided that the required physical networks are present. Available storage includes SAS/iSCSI and EqualLogic high-capacity 3.5" drives. The PoP utilises the OpenStack default back-end drivers and is utilised to deploy VNFs based on the KVM hypervisor[173], although support for Docker containers[174] via vim-emu is also provided.

In addition to the NFVI-PoP, three bare metal servers running ESXi virtualisation[175] software are provided. The ESXi servers are able to provide VMs for other additional key functionalities such as a *Prometheus*[176] monitoring server, various NFVO releases, performance/availability testing tools, penetration testing tools, etc. Additional networking infrastructure includes a Cisco 5500 Series[177]

---

[170] For further information also see: *https://www.openstack.org/software/ocata/*
[171] For more details also see: *http://vault.centos.org/7.4.1708/*
[172] For further information also see: *https://docs.openstack.org/neutron/latest/*
[173] A KVM hypervisor is the virtualisation layer in Kernel-based Virtual Machine (KVM), a free, open source virtualisation architecture for *Linux* distributions. See*, for example*: *https://whatis.techtarget.com/definition/KVM-hypervisor*
[174] A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings. Also see: *https://www.docker.com/resources/what-container*
[175] See: *https://www.vmware.com/products/esxi-and-esx.html*
[176] For more detailed information about Prometheus, see: *https://prometheus.io/*
[177] For further details see: *https://www.cisco.com/c/en/us/products/wireless/5500-series-wireless-controllers/index.html*

Adaptive Security Appliance (integrating firewall, NAT and Intrusion Detection capabilities), a Cisco 2900 Series[178] Integrated Services Router, and two switches, namely an SDN-enabled HPE Aruba 3800[179] with the OpenDaylight[180] controller (version Carbon[181]) and a Dell Switch. The NAT is configured either to be dynamic in order to allow all the hosts to reach internet or public addresses, or static NAT to allow also access to specific services from the inside networks to be reachable outside the firewall.

Furthermore, based on the current local deployment, ORION has provided a simple guide with common considerations and pitfalls with respect to the installation and configuration of an OpenStack VIM on the NFVI, along with the current version of the ETSI-supported OSM orchestrator[182], which is annexed to the current document (Appendix A).

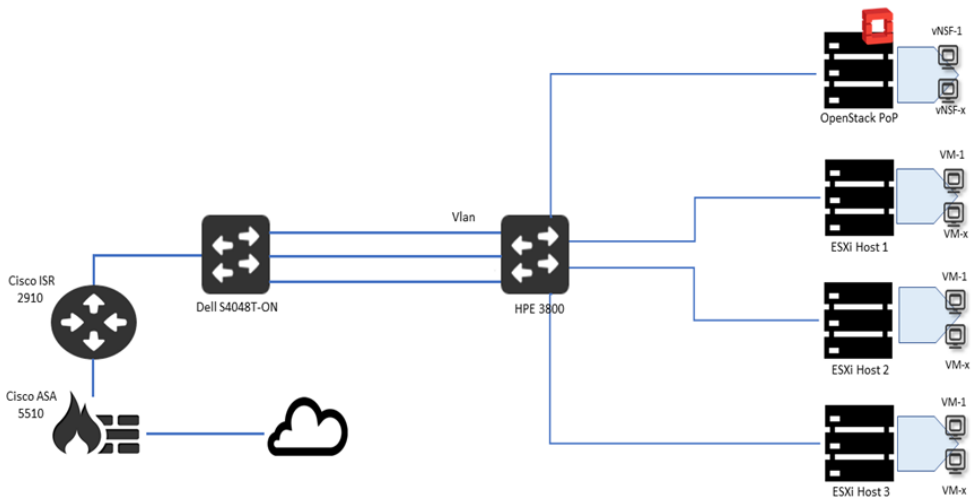


Figure 51: ORION Athens NFVI-PoP (including additional infrastructure).

5.3.2 High-performance VNF development with Enhanced Platform Awareness

When network functions are deployed as VNF-based services, a high level of efficiency is expected, especially in terms of high packet throughput and low latency. However, there are many challenges to overcome, in order to achieve near-native (non-virtualised) performance [47]. For example, packet processing, control plane elements and the VM hypervisor often compete for CPU time. If the packet processing workload cannot be guaranteed uninterrupted CPU time, effects on latency, jitter and packet loss are observed. Hence, the performance of a VM is highly dependent on the underlying infrastructure.

The correct deployment and configuration of the MANO stack (NFVO, VNFM and VIM) & NFVI is of paramount importance for the deployment of highly efficient VNFs. To that extent, Intel has contributed **Enhanced Platform Awareness (EPA)** [48] to the OpenStack community, as a means to

[178] See: *https://www.cisco.com/c/en/us/products/routers/2900-series-integrated-services-routers-isr/index.html*

[179] For further details see: *https://support.hpe.com/hpsc/doc/public/display?docId=emr_na-c03007219*

[180] OpenDaylight (ODL) is a modular open platform for customizing and automating networks of any size and scale. The OpenDaylight Project arose out of the SDN movement, with a clear focus on network programmability. It was designed from the outset as a foundation for commercial solutions that address a variety of use cases in existing network environments. For more details see: *https://www.opendaylight.org/*

[181] For further details also see: *https://www.opendaylight.org/what-we-do/current-release/carbon*

[182] For further information see: *https://osm.etsi.org/*

expose certain hardware characteristics of the NFVI to the OpenStack VIM, which in turn can be used by the deployed VNFs.

With EPA, the VM workloads can be properly matched to those compute hosts that offer them advanced features. Figure 52 illustrates how the deployment process is affected by EPA. The use of EPA ensures that VNFs are instantiated not only based on their state of resource utilisation, but also on the features that the VNF demands. Notable, practical examples are provided in projects such as:

- **5G ESSENCE** [49]**:** where EPA features such as the Data Plane Development Kit (DPDK) [50] and CPU pinning are considered for high-performance VNFs.
- **SHIELD** [51]**:** where EPA security features such as Trusted Platform Module (TPM) are considered for security and attestation.

Table 6 provides some of the NFVI features available to OpenStack on x86 architectures[183], while Table 7 provides performance data for various features published by Intel [52].

Table 6: Enhanced Platform Awareness NFVI features available for OpenStack (as of version Rocky[184] – the list is non-exhaustive).

Compute features	Host CPU feature request
	Real-time kernel and hypervisor
	Single Root I/O Virtualisation (SR-IOV)
	CPU threading
	CPU Pinning
	IO-based NUMA scheduling
	NUMA Awareness
	Huge Pages
	PCIe Passthrough (Example: GPUs utilisation, FPGA, etc.)
	Resource Director Technology (DDR bandwidth allocation, Cache size allocation etc.)
Networking features	Accelerated virtual switch using Data Plan Development Kit (DPDK)
	Accelerated virtual switch using smartNIC
	Data Plane VNF acceleration
Security & Other features	Cryptography (hardware) accelerators such as AES-NI, SIMD/AVX[185], QAT[186]
	Random Data Generation using native instructions such as RDRAND[187]
	Compression Accelerators such as QAT
	Security features such as TPM, SGX, TXT[188] (secure boot & attestation)

[183] The x86 is a family of instruction set architectures based on the Intel 8086 microprocessor and its 8088 variant. The 8086 was introduced in 1978 as a fully 16-bit extension of Intel's 8-bit 8080 microprocessor, with memory segmentation as a solution for addressing more memory than can be covered by a plain 16-bit address. The term "x86" came into being because the names of several successors to Intel's 8086 processor end in "86", including the 80186, 80286, 80386 and 80486 processors. More related information can be found, *for example*, at: *https://en.wikipedia.org/wiki/X86.*

[184] For more details see: *https://releases.openstack.org/rocky/*

[185] See the discussion proposed within the context of the work found at: *https://webcache.googleusercontent.com/search?q=cache:CWzgvMsZ9GcJ:https://www.polyhedron.com/web_images/intel/productbriefs/3a_SIMD.pdf+&cd=3&hl=el&ct=clnk&gl=gr*

[186] For further information see: *https://01.org/intel-quick-assist-technology*

[187] For further relevant information also see, *inter-alia*: *https://en.wikipedia.org/wiki/RdRand*

[188] For informative purposes see the context presented in: *https://en.wikipedia.org/wiki/Trusted_Execution_Technology*

	Hardened Kernel and Hypervisor

Table 7: Performance enhancements per Compute feature, published by Intel.

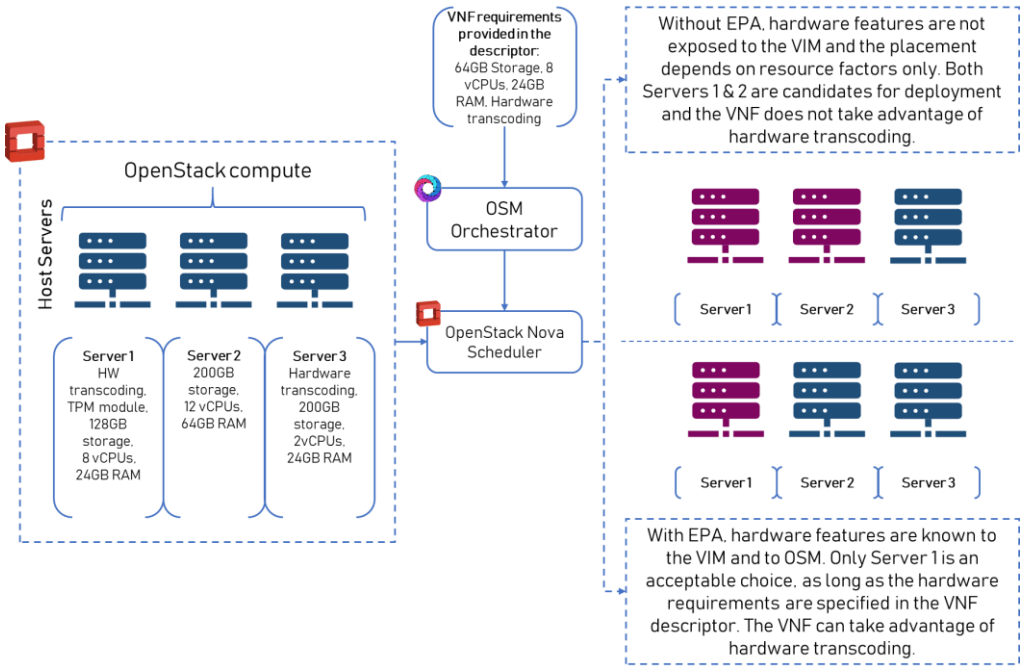
Feature Name	Description	Benefit	Performance Data
Host CPU feature request	Expose host CPU features to OpenStack managed guests	Guest can directly use CPU features instead of emulated CPU features	~20% to ~40% improvement in guest computation
PCI passthrough	Provide direct access to a physical or virtual PCI device	Avoid the latencies introduced by the hypervisor and virtual switching layers	~8% improvement in network throughput
HugePages* support	Use memory pages larger than the standard size	Fewer memory translations requiring fewer cycles	~10% to ~20% improvement in memory access speed
NUMA awareness	Ensures virtual CPUs (vCPU)s executing processes, and the memory used by these processes are on the same NUMA node	Ensures all memory accesses are local to the node and thus do not consume the limited cross-node memory bandwidth, adding latency to memory accesses	~10% improvement in guest processing
IO-based NUMA scheduling	Creates an affinity that associates a VM with the same NUMA nodes as the PCI device passed into the VM	Delivers optimal performance when assigning PCI device to a guest	~25% improvement in network throughput for smaller packets
CPU pinning	Supports the pinning of VMs to physical processors	Avoids scheduling mechanism moving the guest virtual CPUs to other host physical CPU cores, improving performance and determinism	~10 % to ~20% improvement in guest processing
CPU threading policies	Provides control over how guests can use the host hyper thread siblings	More fine-grained deployment of guests on HT[189]-enabled systems	Up to ~50% improvement in guest processing
OVS-DPDK, neutron	An industry standard virtual switch accelerated by DPDK	Accelerated virtual switching	~900% throughput improvement

[189] For more informative details about the case also see: *https://www.intel.com/content/www/us/en/architecture-and-technology/hyper-threading/hyper-threading-technology.html*

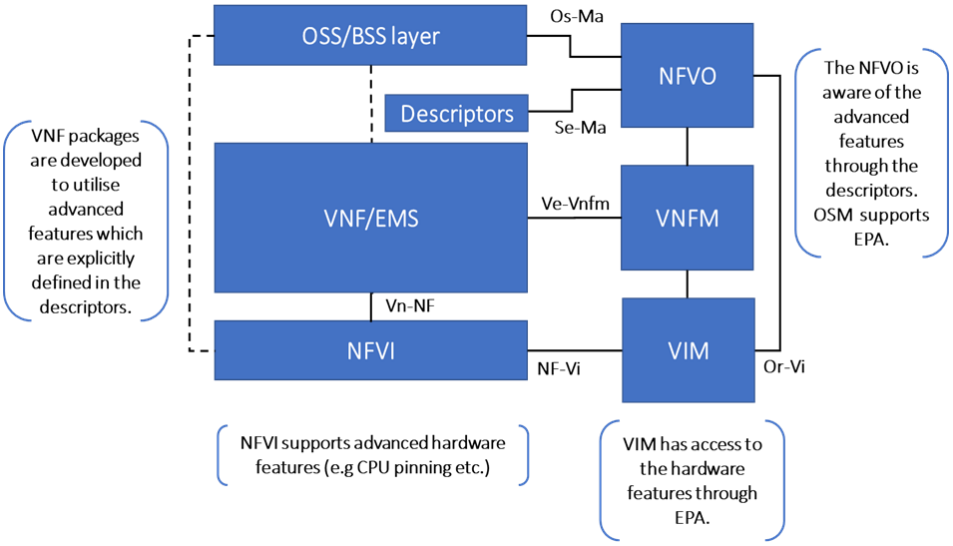
According to a preliminary analysis performed in 5G ESSENCE [49], some important features that can be considered for high-performance VNFs are:

- CPU pinning and CPU threading can dedicate resources to the VM workload, thus alleviating contention for the same resources. This requires configuration of OpenStack and the creation of a VM flavour with pinning enabled, to be used upon instantiation.
- SR-IOV can be used to allow a PCIe device such as a Network Interface Card (NIC) to be accessed by multiple VMs that were previously using virtual NICs. This requires the use of an SR-IOV enabled NIC, as well as the proper configuration of OpenStack.
- DPDK can be used to accelerate switching and improve data plane performance. This requires the configuration of OpenStack networking.

Figure 52 shows how advanced hardware features can be utilised. It is not only necessary for the NFVI to provide those features, but they need to be enabled/supported across the infrastructure.



(a)



(b)

Figure 52: (a) EPA benefits for fine-grained development; (b) Support for advanced features must be configured throughout the infrastructure.

Thus, the implementation of an EPA feature is a complex procedure requiring many configurations that differ per feature and careful selection is advised as incompatibilities, and various problems can arise. Examples of needed configurations for selected features are provided in Table 8 and Table 9.

In general, for a feature to be utilised, the following steps must be taken:

- **Step 1:** Ensure whether the feature is available at the NFVI side: Some features will require BIOS configurations or specialised hardware (such as specialised hardware for TPM security, SR-IOV enabled NICs, etc.). Furthermore, the selected hypervisor technology needs to support the selected feature as well.
- **Step 2:** Ensure that the VIM is properly configured: This entails the creation of flavours (e.g. pinned versus non-pinned to be used during deployment), editing configuration files across OpenStack, ensuring the correct network setup etc.
- **Step 3:** Ensure that the VNF descriptor is correct: The required feature must be part of the VNF descriptor. Otherwise it will not be deployed properly.
- **Step 4:** Ensure that the Orchestrator supports EPA: A multitude of well-known orchestrators support EPA, such as OSM [53] and Kubernetes [54].

Table 8: Example listing the NFVI & VIM configurations necessary (per feature).

	BIOS	Hypervisor	OpenStack
NUMA Topology	X		X
SR-IOV	X	X	X
HyperThreading	X		
Huge Pages		X	
CPU Pinning		X	X

Table 9: Example of the configuration files that need to be edited in OpenStack (per feature).

	nova.conf	ml2_conf.ini	ml2_conf_sriov.ini
NUMA Topology	X		
SR-IOV	X	X	X
CPU Pinning	X		

It is worth noting that EPA is purely an OpenStack contribution. Thus, the vanilla open-source version, as well as the commercial OpenStack versions (such as RedHator MirantisOpenStack[190] and many others), support EPA. Other VIM platforms, such as Amazon's AWS[191], OpenVIM[192], etc., do not support EPA, although this does not preclude that they support selected features (e.g. AWS supports SR-IOV) that require a different process to enable. In any case, it is necessary to carefully select an appropriate VIM that can deliver a required NFVI feature, through careful study of its documentation prior to deployment.

Each feature requires a different process to be enabled in the NFVI; however, there are some common considerations to take into account when planning for an optimal NFVI-VIM deployment:

[190] Also see: *https://www.mirantis.com/blog/mirantis-openstack-7-0-vs-red-hat-rhel-openstack-platform-7/*
[191] See: *https://aws.amazon.com/es/*
[192] See: *https://www.openvim.com/index.html*

(a) Many of these features are only utilised if they are enabled on the BIOS and the kernel of the host system. In the context of the 5G-DRIVE architecture, this is the Linux distribution that is used to deploy the OpenStack private cloud.

(b) An OpenStack enabled NFVI can then implement these features using the Enhanced Platform Awareness framework [48]. After the deployment of OpenStack, the Nova and Neutron services will need to be configured to allow the use of these features. Extensive documentation exists that provides specific instructions per feature.

(c) Special considerations should be made in terms of hardware and software. For example, it is necessary to ensure that the NIC, the virtualisation driver & hypervisor used can also support the required feature.

5.4 The MANO domain

Although MANO is also addressed in Section 4 (and T5.2), this work focuses more on the instantiation and lifecycle management of the VNF-based services, rather than the aspects related to slicing.

5.4.1 General architecture

The Management and Orchestration domain contains all the components necessary to provision the NFVI's resources, instantiate and manage the required Network Services, as detailed in the ETSI MANO specifications [34]. The main functional blocks (also shown in Figure 48) are:

- The **NFV Orchestrator (NFVO)** manages NFVI resources across multiple VIMs and handles lifecycle management.
- The **NFV Manager (NFVM)** passes lifecycle management messages to the running instances. Each VNFM may be assigned to a single instance or multiple ones.
- The **Virtualised Infrastructure Manager (VIM)** controls the NFVI compute, storage and network resources, which usually lie within an operator's domain.

5.4.2 Suggested MANO stack

OpenSource MANO (OSM) [53] is selected as the NFVO. OSM delivers an open source Management and Orchestration (MANO) stack, fully aligned with ETSI NFV Information Models. As an operator-led community, OSM is offering a production-quality open-source MANO stack that meets the requirements of commercial NFV networks. Main features include integrated VIM monitoring, support for multi-VDU VNFs, improved VNF configuration, support for 5G slicing and slice lifetime management, as well as various platform and UI improvements.

Juju [55] is an open-source application modelling tool developed by Canonical Ltd. Its modelling language allows to abstract technical specifics and simplifies management and integration. A juju controller is a service that monitors the events, state and user activities. The use of juju with OSM allows the administrator to control the deployment of a VNF in a cloud-agnostic manner.

OpenStack is considered as the VIM. It is a popular open source solution to deploy and manage public and private clouds. The current (as of March 2019) 18th release, codenamed Rocky [56], addresses many new demands for infrastructures and includes services and improvements to better support NFV, high availability, edge computing, AI/Machine Learning, deployment configurations such as bare metal[193] and FaaS[194], etc. Furthermore, the ability of OpenStack to utilise advanced features on an x86-based NFVI is a key benefit that led to its selection.

5.5 Instantiation and lifecycle management

This subsection describes the process to instantiate and manage/configure a service using the defined MANO stack.

5.5.1 Images, descriptors and records

The Network Service (NS) describes the relationship between VNFs, existing Physical Network Functions (PNFs, if any), and the links required to interconnect them in the NFVI network, or to connect them to other endpoints. A Network Service (NS) is fully characterised by four information elements:

- The Virtualised Network Function (VNF);
- The Physical Network Function (PNF);
- The Virtual Link (VL), and;
- The VNF Forwarding Graph (VNFFG).

Figure 53 illustrates the main information elements defined in ETSI MANO [34]. The information that is included in **descriptors** are deployment templates that provide static information used in the process of on-boarding and instantiating a service. The **records**, *on the other hand*, contain dynamic, run-time information to track the status of the service and the infrastructure.

The **VNF Catalogue** contains the **VNF Descriptors** (VNFDs). The VNFD is a deployment template that describes the deployment and operational behaviour requirements of the VNF. Information typically detailed in the VNFD contains deployment instructions, scaling policies, configuration information and monitoring parameters. Moreover, the VNFD contains connectivity, interface and KPIs requirements that can be used by the NFVO to establish appropriate virtual links between VNF components instances, or between a VNF instance and the endpoint interface to other virtual functions. Furthermore, a separate image repository at the side of the VIM stores the VNF images. A **Network Service Descriptor (NSD)** is a deployment template for a Network Service referencing all other descriptors which describe components that are part of that Network Service.

---

[193] A "bare-metal server" is a computer server that is a "single-tenant physical server". The term is used nowadays to distinguish it from modern forms of virtualisation and cloud-hosting. Bare-metal servers have a single "tenant". They are not shared between customers. Each server may run any amount of work for the customer, or may have multiple simultaneous users, but they are dedicated entirely to the customer who is renting them. Unlike many servers in a data centre, they are not being shared between multiple customers. Bare-metal servers are "physical" servers. Each logical server offered for rental is a distinct physical piece of hardware that is a functional server on its own. They are not virtual servers running in multiple on shared hardware. Also see the discussion proposed in: *https://loige.co/from-bare-metal-to-serverless/*

[194] Function as-a-Service (FaaS) is a category of cloud computing services that provides a platform allowing customers to develop, run, and manage application functionalities without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app. Building an application following this model is one way of achieving a "serverless" architecture and is typically used when building microservices applications.

The **Network Service Catalogue** contains the NS, VNFFG, VL Descriptors (NSD, VNFFGD, VLD respectively). The VNFFGD describes the topology of the NS based on the interconnections. A VNF Forwarding Graph Descriptor (VNFFGD) describes a topology of the Network Service or a portion of the Network Service, by referencing VNFs and PNFs and Virtual Links that connect them. A Virtual Link Descriptor (VLD) describes the resource requirements that are needed for a link between VNFs, PNFs and endpoints of the Network Service, which could be met by various link options that are available in the NFVI. A Physical Network Function Descriptor (PNFD) describes the connectivity, Interface and KPIs requirements of Virtual Links to an attached Physical Network Function. This is needed if a physical device is incorporated in a Network Service to facilitate network evolution. Upon instantiation, operational records are created to represent the newly created instances, based on the descriptor information and various run-time information related to the instances.

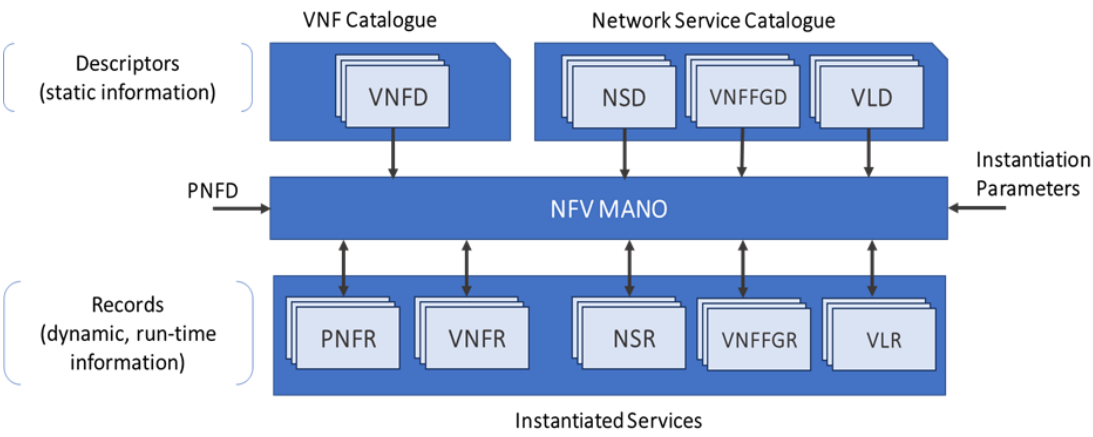


Figure 53: ETSI MANO information elements. Descriptors are provided by the VNF developers, while records are created upon instantiation of a network service.

5.5.2 Instantiation and configuration process

The successful deployment of a VNF relies on the instantiation of the service and its basic configuration. Configuration, in this case, applies to the service and includes basic operational logic and day-1/day-2 configuration, rather than the run-time configuration of its functional elements (which is not handled by the VNFM but rather the EM elements in the ETSI NFV framework). This subsection focuses on instantiation and basic configuration through the VNFM. For a VNF to be instantiated into a service, multiple conditions must be met:

- **The NFVO and VIM must be installed and integrated properly:** OpenStack must be configured to be reachable by OSM. Other special configurations might be required to enable advanced features like Enhanced Platform Awareness (EPA) optimisations[195]. For a detailed description of how to correctly configure OSM and OpenStack, please refer to Appendix A.
- **The VM image must be uploaded on the VIM:** This action can easily be performed via the command line interface or GUI.
- **The VNFD & NSD must be on-boarded (loaded on OSM) and enabled:** This can be performed either through the OSM client CLI or through the OSM web interface. Validation of the descriptor is performed automatically.

Figure 54 and Figure 55 illustrate the on-boarding and enabling process followed by OSM and all ETSI-compliant NFVOs. When the administrator pushes a VNFD to the NFVO, the descriptor is first

[195] For more related information also see: *https://wiki.openstack.org/wiki/Enhanced-platform-awareness-pcie*

validated (i.e. checked for mandatory elements, validated for integrity against security manifests etc.). If successful, the NFVO notifies the Catalogue and makes the VM images available. The descriptor can then be enabled, after additional validation. For example, OSM will check if the descriptor already exists and that it is not marked to be deleted.

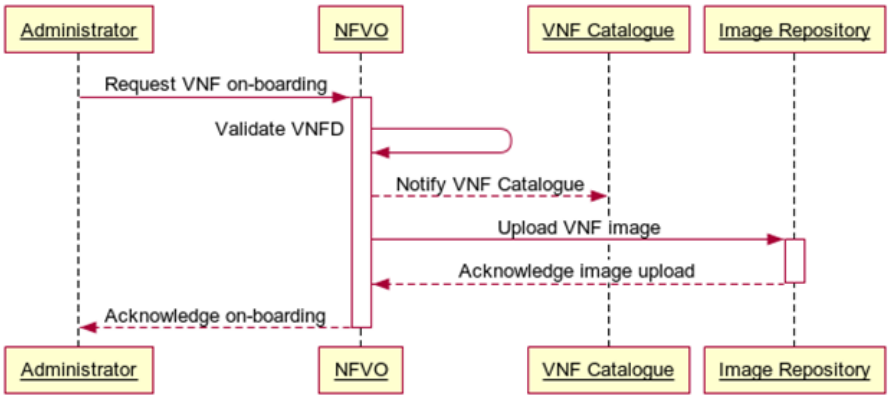


Figure 54: Sequence diagram of the ETSI MANO on-boarding process.

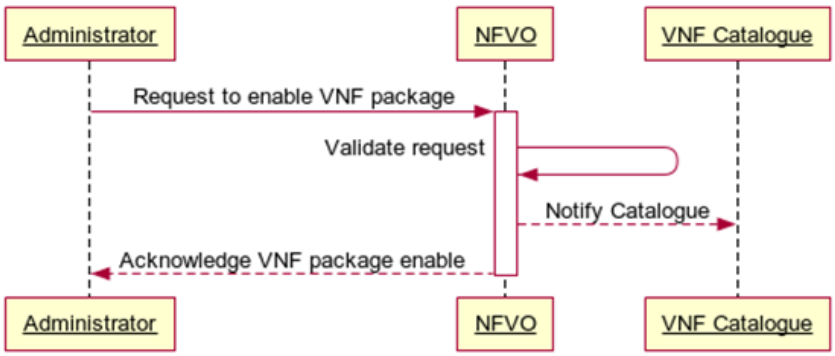


Figure 55: Sequence diagram of the ETSI MANO enabling process.

A similar process is utilised to disable or delete a VNFD from the Catalogue, as each action is first validated. It is worth noting that disabling or deleting a VNF does not affect any running services. It only prevents the VNF from further instantiation. Once the requirements for valid on-boarded VNFD is fulfilled, the administrator can on-board the NSD following a similar process. The NS descriptor that provides additional information on how the VNF should be interconnected within an instantiated service. The OSM administrator is then able to instantiate a service. Multiple services can be instantiated from one VNF.

Figure 56 illustrates the instantiation process. Once the OSM administrator requests the instantiation of service, the following steps are performed:

- The NFVO (e.g. OSM) validates the request by ensuring that there is a valid VNF descriptor that is not disabled or marked to be deleted from the catalogue. Deleting or disabling a VNFD does not affect running services, but such actions prohibit further instantiation.
- The NFVO performs a feasibility check to ensure that the required resources are available from the VIM.

- Resources are allocated, and network interconnections are set up through the NFVO, VIM & NFVM based on the VNFD & NSD information. In OSM, each instantiated VNF corresponds to an LXD container[196] that is launched in OSM to house the Juju proxy charm.
- The basic configuration of the instance is performed by the VNFM: this is based on the available primitives and actions (i.e. defined in the Juju charm). Run-time (functional) configurations can optionally be made through the EM (Elemental Manager)
- The NFVO acknowledges successful instantiation and configuration and informs the administrator. In OSM, the uptime counter is also started.

Once the service has been instantiated, the default VNF primitives can be executed (Table 10). Furthermore, Day-1 and Day-2 configurations and basic operational logic can be defined within a Juju charm as actions.

Table 10: Standard VNF primitives available in OSM (and applied as Juju actions).

Primitive	Description
Config	Change the configuration of a VNF
Upgrade	Perform a VNF Software Upgrade
Start	Start the VNF service
Stop	Stop the VNF service
Restart	Restart the VNF service
Reboot	Reboot the VNF

[196] For more details also see, *for example*: *https://linuxcontainers.org/*

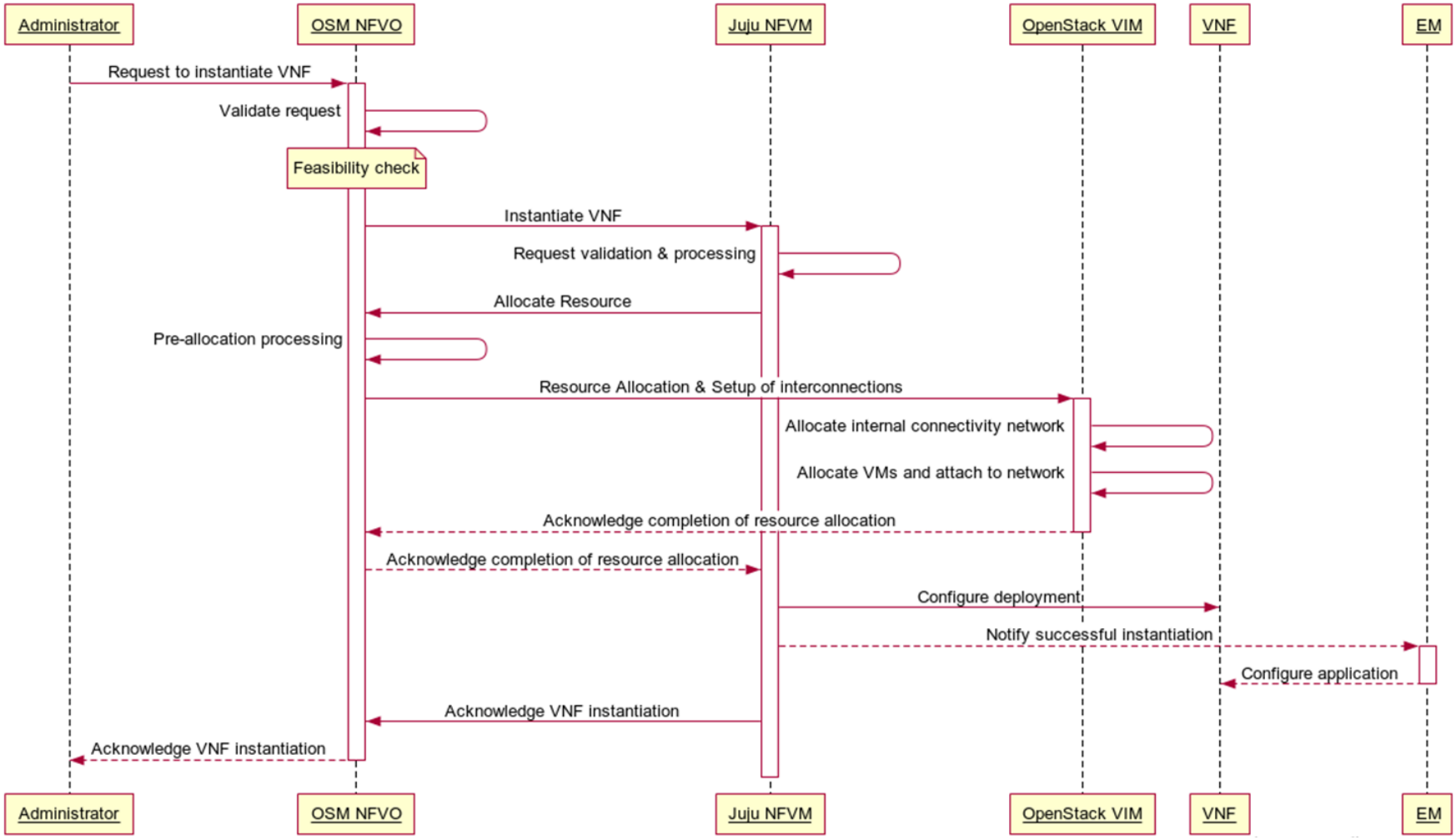


Figure 56: Sequence diagram for the instantiation of a 5G-DRIVE service.

5.5.3 VNF Optimisations with EPA

The OSM Information Model has defined EPA features within the VNF descriptor. Figure 57 shows an excerpt from the Information Model showing the available EPA configurations for the VNFs.

Figure 57: Excerpt from the OSM Information Model for the VNF descriptor, showing definition of EPA features, marking some potential features for the case of the 5G ESSENCE project.

The following subsections provide some key features that can be considered in the 5G ESSENCE deployments.

5.5.3.1 Data Plane Performance

In order to fully exploit the system's resources, both in the network and computational domains, and at the same time enhance and facilitate the implementation of intensive network applications, Intel has developed the **Data Plane Development Kit** (Intel® DPDK) [50]. DPDK comprises of a set of libraries that support efficient implementations of network functions through access to the system's network interface card (NIC). DPDK offers network function developers a set of tools to build high speed data plane applications. DPDK operates in polling mode for packet processing, instead of the default interrupt mode. The polling mode operation adopts the busy-wait technique, continuously checking for state changes in the network interface. This mitigates interruption in packet processing, as it bypasses the kernel, efficiently consuming CPU cycles, which leads to increased packet

throughput. Using DPDK network packet ingress and egress is faster in comparison to the standard Linux kernel network stack as applications are supported in userspace, thus bypassing kernel network stack bottlenecks.

DPDK can be used to accelerate ingress/egress of data plane traffic within the VNF but also to accelerate traffic steering within the OpenStack environment. The use of DPDK-accelerated switching within OpenStack can greatly increase the networking performance and improve scaling of compute hosts (as networking performance is usually the first bottleneck when scaling up the number of compute hosts)[197]. This can be achieved by properly configuring DPDK with the Open Virtual Switch (OVS).
It is important to note that this feature is only supported[198] when using the libvirt compute driver, and the KVM/QEMU hypervisor. In general, the following steps are required:

- Make appropriate settings in BIOS & linux kernel (e.g. enable large or huge pages and the Input Output Memory Management Unit, IOMMU) [57],
- Compile and run OVS & DPDK [58],
- Configure OpenStack Neutron ML2 plugin to utilise OVS-DPDK.
- Include ovs-acceleration in the VNF descriptor.

5.5.3.2 Peripheral Component Interconnect (PCI) device utilisation

Single-Root Input-Output Virtualisation (SR-IOV) [59] is an extension to the Peripheral Component Interconnect Express (PCI Express or PCIe) specification. SR-IOV allows a device (in this case the Network Interface Cards used in 5G ESSENCE Edge DC) to separate access to its resources among various PCIe hardware functions[199] (Physical or Virtual). Each Physical and Virtual Function is assigned a unique PCI Express Requester ID (RID) that allows an I/O memory management unit (IOMMU) to differentiate between different traffic streams and apply memory and interrupt translations between the PF and VFs. SR-IOV enables network traffic to bypass the software switch layer of the virtualisation stack. As a result, the I/O overhead in the software emulation layer is diminished and achieves network performance that is nearly the same performance as in non-virtualised environments. While on SR-IOV a single PCIe device can be virtualised and shared among VMs (in this case, the VMs are the Virtual Deployment Units of a 5G ESSENCE VNF), it is possible to also assign a full physical device to a single VM guest. This is referred to as **PCI Passthrough** [60]. In general, the following steps are required [61] [62]:

- Make appropriate settings in BIOS & linux kernel (e.g. enable IOMMU & SR-IOV).
- Configure OpenStack Neutron to utilise VLAN type driver and sriovnicswitch mechanism driver, as well as define and configure a VLAN range (by editing ml2_conf.ini, ml2_conf_sriov.ini, adding ml2_conf_sriov.ini to the system service and restarting Neutron).
- Configure Nova by adding the PciPassthrough filter on the Nova scheduler, adding a pci_passthrough_whitelist of SRIOV devices, and restarting it.
- For SRIOV, enable the OpenStack Networking SR-IOV agent.

---

[197] The only pitfall is that performance degradation of services that do not support DPDK is to be expected.

[198] Using DPDK in OVS requires the following minimum software versions OVS 2.4, DPDK 2.0, QEMU 2.1.0, libvirt 1.2.13. Multiqueue support is available if the following newer versions are used OVS 2.5, DPDK 2.2, QEMU 2.5, libvirt 1.2.17.

[199] These functions consist of the following types: (a) A PCIe Physical Function (PF). This function is the primary function of the device and advertises the device's SR-IOV capabilities, and; (b) One or more PCIe Virtual Functions (VFs). Each VF is associated with the device's PF. A VF shares one or more physical resources of the device, such as a memory and a network port, with the PF and other VFs on the device.

5.5.3.3 Memory and Central Processing Unit utilisation

The Nova scheduler will, by default, only consider the availability of CPUs when placing a workload. With the advent of Non-Uniform Memory Access **[63]** (NUMA)[200], such behaviour can lead to performance loss, since workloads are not always placed on the same NUMA node. Under a NUMA architecture, optimal performance is achieved when the instantiated VMs utilise the same NUMA node. Furthermore, a deployed VM is often moved by the scheduler, usually for load balancing. Every time a VM is moved, an interruption in CPU time and increase in cache misses are observed, which lead to increase in latency, jitter and packet loss.

Using EPA, it is not only possible to make the scheduler aware of NUMA nodes but also "pin" workloads to specific CPUs so that they will not be moved by the scheduler and maintain NUMA locality. CPU pinning (also known as processor affinity) improves the performance of individuals VMs and reduces the occurrence of cache misses. In general, the following steps are required to enable and use CPU pinning in OpenStack (assuming an x86 architecture):

- Enable NUMA at the compute host BIOS.
- Make sure that the operating system supports NUMA by configuring the linux kernel.
- Make sure that the hypervisors on the compute nodes support NUMA (e.g. libvirt/KVM),
- Configure the Nova Scheduler by adding the appropriate filters (in *nova.conf*).
- Create flavours in OpenStack[201] with CPU pinning enabled (setting *hw_cpu_policy=dedicated*).
- Edit the VNF descriptor to enable CPU pinning and on-board the new descriptor to OSM.

5.6 Integration, testing and assurance

The integration and testing of the network services relate to the integration of the VNFs with the MANO environment, with the NFVI and with other monitoring tools. Furthermore, functional and non-functional (e.g. performance) characteristics need to be ascertained, to ensure that the network services operate within the acceptable limits. Tests that are related to the instantiation and lifecycle management of the services are indicated as mandatory, whereas less critical tests are considered optional. The tests are summarised as follows:

- **Lifecycle Management & Configuration:** The VNFs should be able to be instantiated by OSM, allocated a set of resources from the NFVI. Starting/stopping service, disabling and deleting (decommissioning) a VNF are the minimum requirements. It is noted that while disabling and deleting an on-boarded image, all existing running instances should remain unaffected. Testing also includes basic configurations (day-1 & day-2) through SWA-3 and run-time configurations through SWA-4.
- **Scaling:** These tests relate to scale-in (removing a VNF from a multi-VNF running service), scale-out (adding a VNF to a running service), autoscaling (scaling with automated triggers) or migrating from one service to another one (replacing an NS). These are considered

[200] In traditional symmetric multiprocessing (SMP) architectures, multiple CPUs access the memory through the Front Side Bus (FSB, also known as the "North Bridge"). As the size of the CPU pool grows, the FSB becomes a bottleneck. NUMA divides the memory into multiple memory nodes, "local" to one or more CPUs thus accelerating memory access.

[201] In OSM (release 5) the flavor can be described in the VNFD and Nova can be configured to allow the OSM user to create flavors.

optional since the scenario is not expected to utilise all types of scaling. Autoscaling could be considered in a dynamic scenario with live re-provisioning of existing services. This would require a testbed configuration that supports autoscaling (e.g. OSM Release 5 or newer, scaling parameters defined in the VNF descriptor, related hypervisor technology like KVM, etc.).

- **Service Function Chaining:** This related to the chaining of existing running functions (as opposed to the linking of VNFs within a service with virtual links).
- **Metrics collection & Non-functional testing:** These tests pertain to the non-functional characteristics of the VNFs, e.g. performance, availability, security, etc., and the metrics collected through the VNFs, the VIM and *Prometheus*.
- **Functional testing:** these refer to the testing of the VNF main functionality (i.e. its expected role in a 5G-DRIVE scenario).

End-to-end scenarios can then be created by chaining multiple tests together, at the site of the final integrated testbed.

5.6.1 Integration plan

The following table includes the individual integration tests for the service ecosystem. These tests verify the basic functionalities and performance requirements of 5G-DRIVE VNFs, VIM & MANO. End-to-end tests will be designed in a further stage in the project, to show the ability of a feature-rich infrastructure that supports multiple 5G use case scenarios.

Table 11: Integration plan for T5.3 services with telemetry.

##	ID	Name	Description	Priority[202]	Related Components	Related inputs	Success criteria	KPIs
1	VNF1	VNF deployment	NFVO validates, on-boards & enables a package	M	VNFs, OSM, an OpenStack PoP	Actions from the NFVO admin, VNF image & descriptors	VNF descriptor is on-boarded on Catalogue, or rejected as invalid	Time to complete operation
2	VNF2	VNF instantiation	NFVO instantiates an enabled VNF	M		Actions from the NFVO admin, VNF is on-boarded & enabled	NFVO validates the existence of resources and instantiates a service	Instantiation time
3	VNF3	VNF decommissioning	NFVO disables and deletes an on-boarded VNF	M		Actions from the NFVO admin	VNF descriptor is disabled and removed	Time to complete operation
4	VNF4	NS lifecycle management	The VNF receives LCM messages from the NFVO through the SWA-3 interface and executes the operation	M	A running service, OSM, an OpenStack PoP	Actions from the admin	The NS configuration is applied	
5	VNF5	NS run-time configuration	Configurations are sent to the running service either via direct connection to VM or via SWA-4 interface	M				
6	VNF6	NS migration	Seamless Migration between running instances	O			No faults or interruptions in connectivity	
7	VNF7	NS scaling	Scale out/in a running service by adding or removing a VNF	M				
8	VNF8	NS Autoscaling	Re-provisioning a running instance	O				
9	VNF9	Service Function Chaining	Chain two services together	O	VNFs, OSM, an OpenStack PoP, SDN/RAN controller			

[202] M: Mandatory, O: Optional

##	ID	Name	Description	Priority[202]	Related Components	Related inputs	Success criteria	KPIs
10	VNF10	Export of VNF resource data to Prometheus	The VNF exports resource utilisation data to Prometheus or to OSM Mon container	O	A running instance with node_exporter, a running *Prometheus* server, OSM	None	Prometheus & MON receive streamed data	N/A
11	VNF11	Availability of the network service	The network service runs for 1x, 2x, 4x, 8x the duration of an end-to-end scenario without failures	M	A running instance in an OpenStack PoP	Rally & Os-faults to induce faults and measure KPIs, moongen to generate varied loads	Continuous runtime in variable load conditions	Availability %, MTBF, MTTR
12	VNF12	Network Performance	The network performance of the NS is assessed	M		Shaker, moongen (for network performance), Neutron QoS extension	The NS performs as expected	Latency, packet loss, etc.
13	VNF13	Fault Management	The VNF reports faults in case of failure	O		Rally & OS-faults to induce faults and measure KPIs	The fault is rectified, if the NS has failed catastrophically it is re-deployed	MTTR
14	VNF14	VNF security	Unnecessary services on the VNF are inactive, known vulnerabilities patched, etc.	M		Pentesting/cybersecurity tools, labelled traffic data sets	The NS is hardened against security threats	False positive, false negative detections
15	VNF15	VNF privacy	The VNF provides privacy information, the information is easily discoverable and comprehensible	M	Available as part of specifications and/ or through a UI	Privacy specifications	Privacy information are discoverable	Time to complete a cognitive walkthrough (e.g. time a user trying to find this information)
16	VNF16	Functional testing	Functional testing of the running service	M	A running instance	Specific to the VNF main function		
17	VIM1	Validation of the OpenStack VIM on the NFVI	Main OpenStack services are deployed on a selected infrastructure	M	NFVI infrastructure	OpenStack Tempest is a suite of tests that validate the OpenStack codebase	Depends on the Tempest selected tests	
18	VIM3	BIOS configuration	Selected EPA features (e.g. TPM or NUMA awareness)	O	OpenStack	Enabling required features through the BIOS	N/A	N/A

##	ID	Name	Description	Priority[202]	Related Components	Related inputs	Success criteria	KPIs
			need to be enabled on the BIOS					
19	VIM4	VIM Telemetry	Installation and testing of telemetry services	O	OpenStack, OSM	Ceilometer[203] & Gnocchi[204] services need to be installed and configured to monitor the VIM resource consumption. Sampling rate on OSM must be configured.	OSM receives VIM data on the MON container	N/A
20	MANO1	Testing OSM deployment	OSM deployment must be tested to ensure all required Docker[205]/LXD services[206] are deployed	M	OSM	A fresh installation of OSM	All docker services have been started and Juju is configured properly. Admin default password has been changed	N/A
21	MANO2	OpenStack-OSM configuration	OpenStack must be reachable by the orchestrator	M	OSM, OpenStack	Creation of a user/tenant in OpenStack with the capacity to create flavours, creation of a DHCP enabled management network	OpenStack account is successfully added to OSM	N/A
22	MANO3	SDN controller integration	Configuration of Neutron	O	OpenStack, OSM	If an SDN controller is used, it must be reachable by OpenStack and OSM	N/A	N/A

[203] Also see: *https://docs.openstack.org/ocata/cli-reference/ceilometer.html*

[204] For more related details see, *for example*: *https://wiki.openstack.org/wiki/Gnocchi*

[205] Docker is a set of coupled software as-a-service and platform as-a-service products that use operating-system-level virtualization to develop and deliver software in packages called containers. The software that hosts the containers is called Docker Engine. It was first started in 2013 and is developed by Docker, Inc. The service has both free and premium tiers. Also see: *https://en.wikipedia.org/wiki/Docker_(software)*

[206] Also see, *for example*: *https://linuxhint.com/lxd-vs-docker/*

© 2018 - 2021 5G-DRIVE Consortium Parties

5.6.2 Tools and KPIs

Dynamic placement of functions and service assurance are key elements of T5.3. The common basis for both functionalities is a monitoring system that can gather and stream measurements across the infrastructure. Measurements can then be utilised by the MANO infrastructure for dynamic placement, or for assurance of monitoring/billing, etc. functions that are commonly part of the operational and business support systems (OSS/BSS) layer (shown in Figure 48). Two key factors need to be addressed by the monitoring system:

- Performance & resource utilisation, and;
- Availability of service.

In the first case, measurements can be collected either through a *Prometheus* [64] server, with the addition of node_exporter [65] to the VNFs (Figure 58), via NFVO & VIM monitoring components, or via other tools and frameworks.

Figure 58: Node_exporter use for KPI monitoring from a VM.

Table 12 provides a listing of the KPIs that can be monitored with Node Exporter[207] / *Prometheus*, for *Linux*-based services. *Prometheus* is a useful addition to the overall service ecosystem, as it enables the VNF developer to easily export resource utilisation data as well as hand them to resource-based placement algorithms, to OSM directly, to program alarms and triggers, etc.

Table 12: KPIs collected by Prometheus through use of node_exporter for Linux-based services.

Name	Description	Default status
arp	Exposes ARP statistics from /proc/net/arp.	Enabled
bcache	Exposes bcache statistics from /sys/fs/bcache/.	Enabled
bonding	Exposes the number of configured and active slaves of Linux bonding interfaces.	Enabled
conntrack	Shows conntrack statistics (does nothing if no /proc/sys/net/netfilter/ present).	Enabled
cpu	Exposes CPU statistics	Enabled
cpufreq	Exposes CPU frequency statistics	Enabled
diskstats	Exposes disk I/O statistics.	Enabled
edac	Exposes error detection and correction statistics.	Enabled

[207] See: *https://prometheus.io/docs/guides/node-exporter/*

Name	Description	Default status
entropy	Exposes available entropy.	Enabled
filefd	Exposes file descriptor statistics from /proc/sys/fs/file-nr.	Enabled
filesystem	Exposes filesystem statistics, such as disk space used.	Enabled
hwmon	Expose hardware monitoring and sensor data from /sys/class/hwmon/.	Enabled
infiniband	Exposes network statistics specific to InfiniBand and Intel OmniPath configurations.	Enabled
ipvs	Exposes IPVS status from /proc/net/ip_vs and stats from /proc/net/ip_vs_stats.	Enabled
loadavg	Exposes load average.	Enabled
mdadm	Exposes statistics about devices in /proc/mdstat (does nothing if no /proc/mdstat present).	Enabled
meminfo	Exposes memory statistics.	Enabled
netclass	Exposes network interface info from /sys/class/net/	Enabled
netstat	Exposes network statistics from /proc/net/netstat. This is the same information as netstat -s.	Enabled
nfs	Exposes NFS client statistics from /proc/net/rpc/nfs. This is the same information as nfsstat -c.	Enabled
nfsd	Exposes NFS kernel server statistics from /proc/net/rpc/nfsd. This is the same information as nfsstat -s.	Enabled
sockstat	Exposes various statistics from /proc/net/sockstat.	Enabled
stat	Exposes various statistics from /proc/stat. This includes boot time, forks and interrupts.	Enabled
textfile	Exposes statistics read from local disk. The --collector.textfile.directory flag must be set.	Enabled
time	Exposes the current system time.	Enabled
timex	Exposes selected adjtimex(2) system call stats.	Enabled
uname	Exposes system information as provided by the uname system call.	Enabled
vmstat	Exposes statistics from /proc/vmstat.	Enabled
xfs	Exposes XFS runtime statistics.	Enabled
zfs	Exposes ZFS performance statistics.	Enabled
buddyinfo	Exposes statistics of memory fragments as reported by /proc/buddyinfo.	Disabled
drbd	Exposes Distributed Replicated Block Device statistics (to version 8.4)	Disabled
interrupts	Exposes detailed interrupts statistics.	Disabled

© 2018 - 2021 5G-DRIVE Consortium Parties

Name	Description	Default status
ksmd	Exposes kernel and system statistics from /sys/kernel/mm/ksm.	Disabled
logind	Exposes session counts from logind.	Disabled
meminfo_numa	Exposes memory statistics from /proc/meminfo_numa.	Disabled
mountstats	Exposes filesystem statistics from /proc/self/mountstats. Exposes detailed NFS client statistics.	Disabled
ntp	Exposes local NTP daemon health to check time	Disabled
processes	Exposes aggregate process statistics from /proc.	Disabled
qdisc	Exposes queuing discipline statistics	Disabled
runit	Exposes service status from runit.	Disabled
supervisord	Exposes service status from supervisord.	Disabled
systemd	Exposes service and system status from systemd.	Disabled
tcpstat	Exposes TCP connection status information from /proc/net/tcp and /proc/net/tcp6. (Warning: the current version has potential performance issues in high load situations.)	Disabled
wifi	Exposes WiFi device and station statistics.	Disabled

With respect to the VIM, OpenStack is a very popular choice since it is open-sourced and supports Intel's Enhanced Platform Awareness features. There are multiple tools available to enable testing of OpenStack and Network Service performance and reliability in multitenant scenarios. OpenStack provides extensive documentation on preparing and performing a variety of integration and functional tests on resource consumption, scalability, networking performance, service performance, etc. There are also available tools to inject faults in running instances or the OpenStack infrastructure, emulate service hangs, etc. The following table provides some of the main tools that can be used to measure specified KPIs, both internal and external to OpenStack. These tools can be used to measure specific KPIs.

Table 13: Tools for NFVI & NFV performance & availability tests.

Tool	Description of tool	KPIs	Units	Description of KPI
OpenStack Ceilometer [66]	OpenStack native data collection service, Ceilometer data can be used to provide customer billing, resource tracking, and alarming capabilities	CPU load	MHz	CPU load for each OpenStack service
		RAM consumption	Gb	RAM consumption for each OpenStack service
		Instances amnt	Amount	Max number of instances spawned
		Operation time	msec	Time spent on every instance spawn
Shaker [67]	Data Plane testing tool (needs to be deployed outside OpenStack), can be used to	Latency	ms	The network latency
		TCP bandwidth	Mbits/s	TCP network bandwidth

Tool	Description of tool	KPIs	Units	Description of KPI
	measure tenant networking performance.	UDP bandwidth	packets per sec	Number of UDP packets with 32 bytes payload
		UDP delay jitter	ms	Packet delay variation
		UDP packet loss	%	Percentage of lost UDP packets
		TCP retransmits	packets per sec	Number of retransmitted TCP packets
Rally [68] & OS-Faults [69]	Rally is a generic testing tools to automate & unify verification/testing/profiling on OpenStack nodes. OS-faults can be used to perform destructive actions against OpenStack or Linux-based network services (e.g.: cause or emulate service faults, node faults, power faults etc.)	Service downtime	sec	How long the service was not available, and operations were in error state.
		MTTR	sec	How long it takes to recover service performance after the failure.
		Absolute performance degradation	sec	The mean of difference in operation performance during the recovery period and operation performance when service operates normally.

Availability is of particular importance to the context of the 5G-DRIVE effort. Carrier-grade performance relates to the "five-nines", which translates to 99,999% uptime and "extremely high availability". Moreover, responders only care that the service they rely on works as offered. Work already performed in the context of the 5G ESSENCE features VNFs that monitor the traffic without intrusion in the communications by working with a traffic copy (e.g. an IDS) as well as VNFs that they are inserted in the flow of the traffic (e.g. a firewall (FW)). A fail in the first category may affect the service but not break the responder communications. A fail in the second category could be more critical. For this reason, the degree of reliability and protection must be greater for VNFs that are inserted in the flow of the traffic. For non-critical ICT services, a loss of service lower than 6 minutes annually ("five nines") may be considered adequate. For critical services where it is necessary to avoid any interruptions to the responder communications, "seven nines" or higher might be required. To ensure high availability, the VNF must ensure that failures happen rarely and that recovery times are reasonable. Thus, the Mean-Time-Between-Failures (MTBF) and Mean-Time-To-Repair (MTTR) can be used to estimate the expected availability of the service:

$$\text{Availability } (\%) = \frac{\text{MTBF}*100}{\text{MTBF}+\text{MTTR}} \qquad \text{Equation 10.}$$

Table 14: Availability characteristics.

Availability %	Downtime per year	Downtime per month	Downtime per week	Downtime per day
99.999% ("five nines")	5.26 minutes	26.30 seconds	6.05 seconds	864.00 milliseconds

99.9999% (“six nines”)	31.56 seconds	2.63 seconds	604.80 milliseconds	86.40 milliseconds
99.99999% (“seven nines”)	3.16 seconds	262.98 milliseconds	60.48 milliseconds	8.64 milliseconds

Apart from infrastructure faults, the common threat to availability comes from cyber-attacks. This work is considered out-of-scope for the current MS9 report but will be reported in future documents and milestone reports in WP5. Some initial considerations are provided in the following sections.

5.7 Security and privacy aspects

As defined in the 5G-DRIVE DoA, a substantial part of the work undertaken by WP5 relates to the identification and analysis of future security challenges in communications channels of IoV within the 5G complex ecosystem. The 5G-DRIVE will address meticulously a greater concern for connected vehicles in 5G future vehicular networks: secure and privacy-friendly communications.

5.7.1 Security in 5G V2X

To secure a vehicle adequately, the security must be considered from the design and at levels and especially at the networking and the MEC level. In the context of SDN-based 5G interfaces, potential attacks are likely to increase. For this reason, the University of Luxembourg (Uni.lu) partner of the 5G-DRIVE consortium is studying the location privacy protection against an external global passive adversary. This attacker aims to track the target vehicle by eavesdropping all communications of any vehicle within a region of interest. To mitigate this attack, Uni.lu is proposing a context-aware SDN-based pseudonym changing strategy. This strategy uses SDN controllers as the strategy coordinators and relies on them to change the security parameters of pseudonym changing strategy. This proposed strategy supports both infrastructure and infrastructure less vehicular zones. The pseudonym changing strategy considers the context where vehicles are evolving. The context is defined by the pseudonym changing strategy security parameters that are forwarded by the global SDN controller. It mainly includes the threshold of privacy (calculated based on the average of the preferred levels of privacy protection), the number of required vehicles (which depends on the density of vehicles), and the strategy timeout.

© 2018 - 2021 5G-DRIVE Consortium Parties

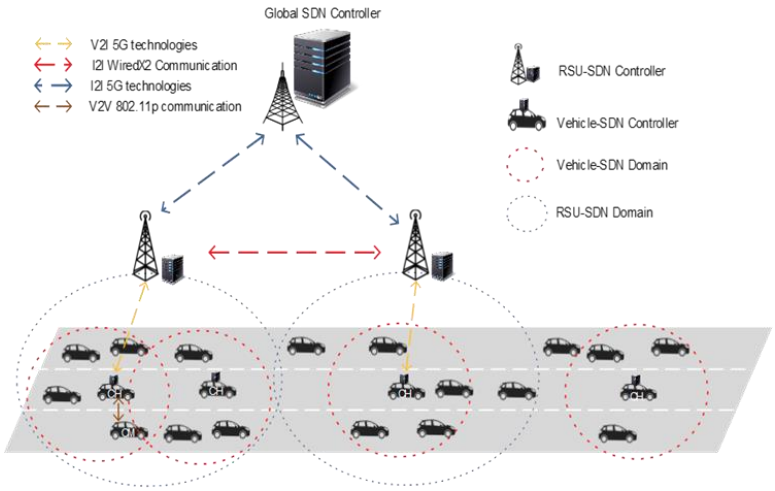


Figure 59: Software defined vehicular network architecture

We consider a hierarchical SDN vehicular network architecture (see Figure 59). We also assume that vehicles periodically form clusters that change over time. The clustering helps to reduce radio interference and overhead and to provide better support for density and mobility. In addition, as the cluster head (CH) will play the role of a local SDN controller, the clustering method implemented by the global SDN controller should then ensure the maximum stability of the CH to minimize how often the CH, and hence the local SDN controller, is changed. As illustrated in Figure 59, there are three SDN control levels in our architecture. The first is a local SDN that operates at the level of the cluster and which coordinates the PCS inside the cluster. This local SDN controller is called Vehicular-SDN Controller (VSDNC), and its cluster is called the VSDNC domain. The second level of the control is at Road-Side Units (RSUs). Each RSU includes an SDN controller and has its own control domain. The RSU-SDN Controller (RSDNC) coordinates different vehicular clusters that are within its range. Finally, the third level is the global SDN controller, which has global knowledge about the vehicular network. The remainder of the vehicles belong to the forwarding plane. Each vehicle is equipped with an IEEE 802.11p[208] interface to communicate with other vehicles and with RSUs. An SDN controller and an SDN agent are also installed in each vehicle. While the SDN agent should be always activated, the SDN controller is initially deactivated and will only be activated when the vehicle becomes a CH and deactivated again if the vehicle reverts to being a cluster member. The internal clocks of vehicles are synchronized using GPS signals. Each RSU is also equipped with two interfaces: wired link to communicate with the neighbouring RSUs, and an LTE interface to communicate with the global SDN controller. This latter is hosted at a distant location. We assume that the communication links between the VSDNC, the vehicles, and between the three types of SDN controllers are secured.

[208] The IEEE 802.11p is an approved amendment to the IEEE 802.11 standard to add wireless access in vehicular environments (WAVE), a vehicular communication system. It defines enhancements to 802.11 (the basis of products marketed as Wi-Fi) required to support Intelligent Transportation Systems (ITS) applications. This includes data exchange between high-speed vehicles and between the vehicles and the roadside infrastructure, so called V2X communication, in the licensed ITS band of 5.9 GHz (5.85-5.925 GHz). IEEE 1609 is a higher layer standard based on the IEEE 802.11p. It is also the basis of a European standard for vehicular communication known as ETSI ITS-G5. For more informative details also see, *for example*: *https://en.wikipedia.org/wiki/IEEE_802.11p*

5.7.1.1 Adversary Model

We assume an external passive adversary who aims to track the trajectory of target vehicle by eavesdropping all communications of any vehicle within a region of interest. We only consider syntactic attacks and not semantic attacks. The adversary is not able to perform tracking using cameras, because the cost of global eavesdropping with cameras is much higher than radio-based eavesdropping. Consequently, camera-based global eavesdropping is beyond the scope of this work. The power of the adversary has a direct impact on the location privacy level of vehicles. If the adversary is strong then the level of location privacy may rapidly decrease following the PCS. The location privacy level can be expressed as function of the power of the adversary.

5.7.1.2 SDN-BASED Pseudonym Changing Strategy

In the following, we describe the main steps of our proposed SDN-based PCS comprising of: (i) the installation of the security parameters of the PCS; (ii) the local SDN monitoring and the PCP, and finally; (iii) the dynamic changing of the PCS security parameters.

- **Installation of PCS security parameters**

  This step installs the PCS's security parameters in each SDN-controller and each vehicle. First, the global SDN controller installs the default PCS security parameters. However, these parameters will be dynamically updated by the SDN controller to adapt to security changes that have occurred. After the creation of the vehicular clusters, the global SDN controller sends the security parameters of the PCS to RSDNCs. Each RSDNC installs these security parameters as soon as it receives them. It will then send back an acknowledgment to the global SDN controller and forwards these security parameters to each VSDNC within its transmission rage. However, in the case of the infrastructure-less scenario, the VSDNC may be outside the range of the RSDNC. For this reason, we assume that a set of default PCS security parameters are already installed at the VSDNC. These parameters will be updated as soon as the VSDNC is in range of a RSDNC. Each VSDNC will then install the PCS security parameters and send back an acknowledgment to RSDNC. Finally, it sends the necessary PCS security parameters to each cluster member and starts the monitoring step

  The following security parameters are considered by the proposed PCS:

  - The threshold of privacy (α): the threshold below which the vehicle should change its pseudonym.
  - The frequency of changing of pseudonyms (β): defines the number of pseudonyms that will be used each hour.
  - The number of required vehicles (γ): defines the number of candidate vehicles required to initiate the PCP.
  - The strategy timeout (δ): defines the duration after which the PCP should be initiated.

- **Monitoring and pseudonym-changing process**

  In this step, the local SDN controllers (VSDNCs) start monitoring the clusters to check whether the conditions are met to initiate the PCP. These conditions are defined by the previously installed PCS security parameters. The VSDNCs also inform the global SDN controller about the result of the PCP. This information will be used to update the security parameters, as described in the next paragraph. Moreover, each SDN-agent belonging to the forwarding plane, periodically sends an update to its VSDNC. These updates generally include the mobility parameters (position, speed, and acceleration) of vehicles and their current privacy level. These updates are used by VSDNCs to select the vehicles that can participate in

the next the PCP. We exploit the safety messages to send these updates. The mobility parameters are by default sent within the beacons and the extension fields are used to send the current privacy level of vehicles.

A vehicle v_i is selected to participate in the next process of changing of pseudonym, only if the vehicle meets a specific context. The context is defined by the PCS security parameters that are forwarded by the global SDN controller. It principally includes the threshold of privacy, the number of required vehicles, and the strategy timeout.

- **Update of security parameters**

  The SDN controllers operating at the three levels of the control plane exchange information to ensure the efficiency of the applied PCS. Each VSDNC reports information to its regional domain controller (RSDNC) in order to keep track of vehicles changing their clusters. In addition, VSDNCs report information about the efficiency of the applied PCS to the global SDN controller via the RSDNs. The purpose of this information is to tune the PCS security parameters in order to obtain better location privacy protection. The PCS security parameters can be tuned as follows:

  - The frequency of changing of pseudonyms (β): a higher frequency value has a positive impact on the location privacy. However, higher change frequency can have negative impacts on the application performance and will increase the number of pseudonyms used.
  - The threshold of privacy (α): this parameter could also be tuned by the global SDN controller according to referred levels of privacy protection, which are provided by users. For instance, this parameter could regularly be calculated based on the average preferred levels of privacy protection. This parameter could also be impacted by the power of the adversary. Indeed, the recommended threshold of privacy should dynamically be increased or decreased according to the power of the adversary.
  - The number of required vehicles (γ): a high number of vehicles changing their pseudonym together has a positive impact on location privacy protection. However, as long as, the decision to initiate a PCS process depends on obtaining a required number of vehicles, the PCS may not perform well if this parameter is not well tuned.
  - The strategy timeout (δ): this parameter is closely related to γ parameter. It ensures that the PCS is executed even if the number of required vehicles is not achieved. This parameter should be tuned to prevent executing unnecessary iterations of the PCS.

5.7.2 Privacy in 5G V2X

In parallel, the analysis and identification of personal data protection challenges related to the 5G-DRIVE vehicular network is of key relevance to the success of the project, and to this end, a dedicated task has been specified to research and analyse the risks related to shared environments, where the same infrastructure is shared among different actors. This phenomenon can be due to either technological particularities (network slicing, for example) or to the varying reach of data privacy regulations across countries.

Initial work has been directed towards the identification of the key privacy problems that are relevant to V2X networks, along with the clarification of the state-of-the-art for privacy-enhancing technologies that might contribute to the solution of these problems. Towards this end, and in alignment with the defined actions of the DoA, an initial work towards the EU-China cross-

examination of relevant privacy and data protection regulations and standards has been carried out (and will continue to evolve throughout the project).

Table 15: EU and China regulatory framework.

Europe	Privacy	General Data Protection Regulation: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)[209] (Text with EEA relevance)
		ePrivacy Directive: Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)[210]
		Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications)[211].
	Open Internet	Open Internet Regulation: Regulation (EU) 2015/2120 of the European Parliament and of the Council of 25 November 2015 laying down measures concerning open internet access and amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services and Regulation (EU) No 531/2012 on roaming on public mobile communications networks within the Union (Text with EEA relevance)[212].
	Security & Law Enforcement	Data protection in criminal investigations: Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA[213].
		Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union[214].
	Non-discrimination	Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and

[209] Text available at: *http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679*
[210] Text available at: *https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32002L0058*
[211] Text available at: *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017PC0010*
[212] Text available at: *https://publications.europa.eu/en/publication-detail/-/publication/8fdf5d08-93fc-11e5-983e-01aa75ed71a1/language-en*
[213] Text available at: *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2016.119.01.0089.01.ENG*
[214] Text available at: *https://eur-lex.europa.eu/eli/dir/2016/1148/oj*

	and Human Rights	occupation[215].
		European Charter of Fundamental Human Rights, esp. Article 8(1) on the protection of personal data[216].
		Treaty of Amsterdam[217] (1997/1999 establishing the protected grounds against discrimination) & Treaty of Lisbon[218] [63] (2007/2009 making the ECHR Bill of Rights legally binding)
		Council of Europe recommendations on profiling: Recommendation CM/Rec(2010)13 of the Committee of Ministers to member states on the protection of individuals with regard to automatic processing of personal data in the context of profiling[219].
China (current description to be evolved)	Personal Data Protection	Personal Information Security Standard GB/T 35273-2017[220]
	Security	PRC Cybersecurity Law[221]
	Consumer Rights	PRC Law on the Protection of Consumer Rights and Interests[222]
	Telecom sector	PRC Telecom Regulation[223]

Additionally, initial work on privacy by design and by default solutions for V2X has been carried out (Uni.lu), which will feed directly the work done by other T5.4 partners. In this context, since location privacy is of paramount importance for drivers, Uni.lu is designing Vehicular Location Privacy Zones (VLPZs) that provide for location privacy protection in the IoVs [70]. A VLPZ is defined as a specific zone managed by trusted regional authorities like municipalities or directly by the country transportation department. The VLPZ aims to increase the location privacy protection level of vehicles within the considered area by providing an effective pseudonym changing. It is crucial for transportation authorities to place these VLPZs within the smart city. For this reason, Uni.lu is proposing an optimal and dynamic placement of VLPZs taking reducing deployment and transportation costs as an objective function. The approach uses a genetic algorithm for optimal placement of VLPZs.

In addition, an initial theoretical examination of the impact of network slicing (with the involvement of several 5G-DRIVE partners such as MI, Uni.lu and OTE) towards personal data protection, is currently underway (MI), which will consider the security work detailed supra and privacy-specific

[215] Text available at: *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32000L0078*
[216] Text available at: *https://fra.europa.eu/en/charterpedia/article/8-protection-personal-data*
[217] See, *for example*: *https://webcache.googleusercontent.com/search?q=cache:_jvz4lBPn1AJ:https://europa.eu/european-union/sites/europaeu/files/docs/body/treaty_of_amsterdam_en.pdf+&cd=1&hl=el&ct=clnk&gl=gr*
[218] See, *for example*: *https://www.robert-schuman.eu/en/european-issues/0173-fundamental-rights-protection-in-the-eu-post-lisbon-treaty*
[219] Text available at: *https://www.coe.int/en/web/data-protection/legal-instruments*
[220] See, *for example*: *https://www.insideprivacy.com/tag/chinese-data-protection-standard-gb-t-35273-2017/*
[221] See, *for example*: *https://en.wikipedia.org/wiki/China_Internet_Security_Law*
[222] See, for example: *http://www.lehmanlaw.com/resource-centre/laws-and-regulations/consumer-protection/law-of-the-peoples-republic-of-china-on-protection-of-the-rights-and-interests-of-the-consumers-1994.html*
[223] See, *for example*: *http://www.zhonglun.com/Content/2017/08-16/1841302098.html*

security requirements identified in the examination of the regulations, standards and recommendations carried out supra.

Finally, the task seeks to define a set of legal compliance specifications for VNFs (with the involvement of several 5G-DRIVE partners such as MI, OTE and Uni.lu), considering the lessons learned from the 5G-DRIVE use-cases and other WP2 outputs. While work on this has been delayed by the initial phases of the project and the lack of use-case specifications, a general approach has been envisioned, under which MI will seek to perform a techno-legal analysis based on the technical inputs provided by University of Luxemburg and OTE, along with the outputs of other 5G-DRIVE WPs as necessary.

The final version of D5.3 will provide more information on security and privacy in 5G and V2X. A preliminary version will be submitted at M18.

© 2018 - 2021 5G-DRIVE Consortium Parties

6 Conclusions

6.1 Final remarks

So far, the work performed in T5.1 has shown that in an orthogonal frequency division multiplexing (OFDM) based massive multiple-input multiple-output (MIMO) system, when analogue beamforming is involved and the ratio of the channel bandwidth to central frequency is relatively large, the beam formed at a frequency other than the central frequency will suffer from the phenomenon known as beam squint: it will point to a direction different from the direction targeted by the central frequency. The beam squint will cause a non-flat channel condition and will affect the system performance significantly. Furthermore, the Transport Network evolution necessary for 5G and beyond mobile systems has been outlined. This work has focussed on two areas. The first of these is on converged, Ethernet-based fronthaul, where lower-layer split (LLS), with adaptations for low-latency transport, and higher-layer split (HLS) interfaces can be transported over the same infrastructure. Measurements of performance using hardware probes in the network can enable dynamic control and adaptation using SDN controllers that are responsive to the Class of Service requirements of different flows.

Deep learning solutions for multiuser single-input multiple-output (MU-SIMO) coherent detection have also been studied. According to the manner of utilizing the channel state information at the receiver side (CSIR), deep learning solutions are divided into two categories: channel equalisation-and-learning and direct learning. It has been shown that direct-learning solutions outperform the equalisation-and-learning solutions due to their better exploitation of the sequence-detection gain. On the other hand, the direct learning solutions are not scalable to the size of MU-SIMO networks, as current DNN architectures cannot efficiently handle many co-channel interferences. This observation is motivating further research towards scalable direct-learning based multiuser-MIMO detection.

The exploitation of DU/CU structure in 5G RAN for improved resource allocation, and load-balance is studied within T5.1. The hierarchical control method is proposed for hybrid use association, to achieve the load balance in a multi-tier network and to reduce the inter-cell interference. In this method, the CU makes the centralised decision to adjust the handover parameter of individual DUs, based on simple and abstracted parameters received from the DUs. The DU, which is the 5G base station, makes its own user association decision based on the tuned handover parameter. The advantage of this solution is the simple and scalable centralised approach only based on simple parameters from DUs.

With respect to slicing (T5.2), the basis of the work is founded on the definition of KPIs for network slices. Such set of KPIs is of premium importance for operators in order to compare solutions of different vendors for the verification of proper operations of the installed systems (also as a part of SLAs with business customers) and for trials of newly developed solutions. The goal was to define a minimal but representative set of KPIs that will describe the network slicing impact on the behaviour of the sliced solutions. However, the 5G-DRIVE project considers that more work is needed for network slicing KPIs evaluation and estimation. For example, the network slice metrics that include the number and size of footprints of all VNFs that compose the slice, number of slice links, number of operations concerning slice configuration can be used for the estimation of slice deployment time. As noted, some KPIs are tightly coupled with MANO orchestration and should be defined as a part of MANO, although so far this is not the case.

With respect to novel service development (T5.3), 5G-DRIVE has initiated work on defining the required software stack (OSM, OpenStack, Juju, virtualisation components, etc.). At this stage, work is completed in individual local testbeds owned by WP5 partners. This document provides a description of related standards, performance features, related KPIs and tools to monitor the status

of the infrastructure and services. This is considered a key development for T5.3, as the monitoring framework is a common ground on which functionalities such as billing, fault management, assurance, dynamic placement etc. are built. Specifically, according to the ETSI NFV standard, these functionalities are mapped either to the MANO or the OSS/BSS layer, although the collection of measurements must apply to the entire infrastructure. Another important aspect is the configuration of hardware awareness. This is essential as it provides advanced placement of VNFs in order to benefit from required hardware features.

6.2 Future work and important considerations

Within the first months of the project, WP5 has dedicated extensive effort to identify the practical issues that hinder the development of novel 5G infrastructures and services and address them by using pragmatic, technical approaches. It is necessary, *however*, to continue efforts in order to provide further improvements.

WP5 has already provided a resource allocation-based solution proposed to compensate for the effect of beam squint. It further aims to focus on improving the system model and analysing the problem and solution for a multi-user system. A necessary next step is also to integrate SDN control with network orchestration, management and control functions. The second focus of the work has been on developing flexible, digitally processed subcarrier multiplexing for analogue transport of antenna signals, in an LLS fronthaul. The technique has been demonstrated experimentally and in benchmarked simulations. In future work, the definitions for the flexible control of the multiplexes is planned, enabling interoperability with virtualised infrastructure managers, and SDN and orchestration functions. Not only the user association but also the spectrum allocation will be considered to use the proposed hierarchical control methods, for the load balance and network capacity improvement.

Going forward, WP5 will better define the performance and reliability requirements per slice and use-case type (eMBB, mMTC, URLLC) and focus on mobility management and on OSS/BSS applications. OSS/BSS includes fault management, billing, accounting etc. Furthermore, the deployment of control plane applications as services and related optimisations will be taken into account for future work.

6.3 Contributions to WP3 & WP4

Deliverable D5.2 will present a detailed plan for WP5 to contribute to WP3 and WP4 developments. This section discusses some initial plans from individual partners to align WP5 innovation with WP3/4.

6.3.1 Network slicing

Orange plans to deploy the network slicing KPIs described in this deliverable, by the use of the Open Source MANO orchestrator (OSM). The initial plan is to evaluate the implement these KPIs in an isolated laboratory environment with artificially created network slice templates (blueprints). Later on Orange will analyse how to use the software, which will be a part of the OSS/BSS according to MANO, in 5G-DRIVE testbed. It is expected that the list of KPIs and the ways of calculation may be further modified, based on the gained experience.

6.3.2 Service evolution and testing

ORION intends to deploy and test WP3 services (such as video caching, reverse proxy) as well as T5.3 services (deep packet inspection) on its dedicated WP5 testbed. This will enable WP5 monitoring

tools, scaling policies etc. to be used with the WP3 VNFs. ORION intends to gather measurements not only regarding performance (WP3), but also availability and reliability (WP5). Multiple NFVI EPA features specific to the eMBB use case will be utilised so as to measure their actual impact on performance. This will enable to correlate the WP3 performance measurements with the effects of NFVI hardware acceleration and with availability stress tests (WP5). Furthermore, ORION will examine how to integrate WP3 VNFs with the WP5 Apache Spot[224] cybersecurity monitoring framework[225], so as to perform traffic sampling and send samples directly to be ingested by the Machine Learning algorithm. The training of the cybersecurity algorithm with use-case specific traffic captures will be examined in the future. The rationale for the integration of the ORION WP3/5 components on a single testbed allows for cybersecurity and fault injection testing in a controlled environment without risk to the trial sites.

6.3.3 Channel modelling for MIMO systems

Currently, according to the results obtained from trials, in the MIMO systems, especially under the environment with mobility[226], the channel may not be stationary. Therefore, traditional channel model based on the assumption of stationary channel may not be suitable. As already mentioned in Section 2.2, UKent is now investigating to use two approaches to model non-stationary channels: (i) Dent model, also known as the modified Jake's model, and; (ii channel transformation into visual regions (VRs) [3] and non-visual regions (nVRs).

In 5G-DRIVE, under WP3, the multi-antenna performance tests examine the performance enhancement, measured through SS-RSRP, SS-SINR and CSI-SINR and UL/DL throughput, obtainable by multiple narrow beams in comparison to wider beams. In WP5, these two models will be simulated and recommended to WP3, and then compared and tested with the results obtained from the trials in order to obtain a more suitable channel model.

6.3.4 Evolved Ethernet network: latency and SDN processing

Latency issues and SDN processing for the fronthaul were discussed in Sections 3.2 and 3.4 respectively. Currently, SDN measurements have been carried out for a limited number of use cases and at 4G data rates. UKent plans to contribute to WP3 outcomes by cooperating with VTT in latency measurements within VTT's fronthaul network. Similarly, the SDN algorithms that have been developed can be tested by incorporating an SDN switch within VTT's network. The aim will be to investigate SDN traffic steering and load balancing in a "real" 5G network and demonstrate the ability of pure Layer-2 SDN processing to reduce latency and latency variation of user plane and control plane traffic. Somewhat more tentatively, latency measurements and SDN control can be incorporated into one of the trial sites, provided that an Ethernet fronthaul network deployment occurs.

[224] See, for example: *https://spot.incubator.apache.org/*
[225] For further details see: *https://www.cloudera.com/solutions/gallery/intel-apache-spot.html*
[226] 5GDRIVE Project, Deliverable D3.2: "Joint Specification for eMBB Trial"

7 References

[1] H. Steyskal, "Digital beamforming-an emerging technology", in *IEEE Military Communications Conference 1988*, San Diego, CA, USA, 23-26 Oct. 1988.

[2] Z. Wang, L. Cheng, J. Wang and G. Yue, "Digital compensation wideband analog beamforming for millimetre-wave communication", in *IEEE 87th Vehicular Technology Conference (VTC Spring)*, Porto, Portugal, 3-6 June 2018.

[3] S. K. Garakoui, E. A. M. Klumperink, B. Nauta and F. E. v. Vliet, "Phased-array antenna beam squinting related to frequency dependency of delay circuits", in *IEEE 8th European Radar Conference 2011*, Manchester, UK, 12-14, Oct. 2011.

[4] F. Rusek, P. D. B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays", *IEEE Signal Processing Magazine,* vol. 30, no. 1, pp. 40-60, 2013.

[5] S. Xue, Y. Ma, A. Li, N. Yi and R. Tafazolli, "On unsupervised deep learning solutions for coherent MU-SIMO detection in fading channels", in *IEEE ICC'19*, 2019.

[6] A. Shojaeifard, K. Hamdi, E. Alsusa, D. So and W. K.K., "On the design of irregular HetNets with flow-level traffic dynamics", in *in Proc. IEEE Veh. Tech. Conf.*, Montreal, Canada, 2016.

[7] K. Stamatiou and H. M., "Traffic management in random cellular networks", in *Inf. Theory and Applications*, San Diego, CA, USA, 2014.

[8] H. Klessig, D. Ohmann, A. Fehske and G. Fettweis, "A performance evaluation framework for interference-coupled cellular data networks", *IEEE Transactions on Wireless Communications,* vol. 15, no. 2, pp. 938-950, February 2016.

[9] H. Kim, G. Veciana, Y. X. and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks", *IEEE Transactions on Networking,* vol. 20, no. 1, pp. 177-190, February 2012.

[10] A. Duda, "Diffusion approximation for time-dependent queueing systems", *IEEE Sel. Areas Commun.,* vol. 4, no. 6, pp. 905-918, 1986.

[11] A. Antonnen and A. Mämmelä, "Interruption probability of wireless video streaming with limited video lengths", *IEEE Trans. Multimedia,* vol. 16, no. 4, pp. 1176-1180, June 2014.

[12] H. Tang, J. Peng, P. Hong and K. Xue, "Offloading performance of range expansion in picocell networks", *IEEE Wireless Commun. Letters,* vol. 2, no. 5, pp. 551-514, October 2013.

[13] H. Du, Y. Zhou, L. Tian, X. Wang, Z. Pan, J. Shi and Y. Yuan, "A load fairness aware cell association for centralized heterogeneous networks", in *Proc. Int. Conf. Commun.*, London, UK, 2015.

[14] 3GPP, "Study on New Radio Access Technology: Radio Access Architecture and Interfaces, 3GPP TR 38.801, v14.0.0", April 2017.

[15] The iCirrus Consortium, "Deliverable D3.4, Updated Low-Cost Energy Efficient Fronthaul Architecture.", [Online]. Available: http://www.icirrus-5gnet.eu/d3-4-updated-low-cost-energy-efficient-fronthaul-architecture/ . [Accessed April 2019].

[16] G. Birring, P. Assimakopoulos and G. N.J., " An Ethernet-Based Fronthaul Implementation with MAC/PHY Split LTE Processing.", in *IEEE 2017 Global Communications Conference, GLOBECOM 2017*, 2017.

[17] P. Assimakopoulos, J. Zou, K. Habel, J.-P. Elbers, V. Jungnickel and N. Gomes, "A Converged Evolved Ethernet Fronthaul for the 5G Era", *IEEE Journal on Selected Areas in Communications, vol. 36, no. 11, pp. 2528-2537, Nov. 2018,* vol. 36, no. 11, pp. 2528-2537, November 2018.

[18] S. Noor and P. Assimakopoulos, "A Flexible Subcarrier Multiplexing System with Analog Transport and Digital Processing for 5G (and beyond) Fronthaul", *submitted to OSA/IEEE J. Lightwave Technology,* 2019.

[19] N. Gomes, P. Sehier, H. Thomas, P. Chanclou, B. Li, D. Muench, P. Assimakopoulos, S. Dixit and V. Jungnickel, "Boosting 5G Through Ethernet: How Evolved Fronthaul Can Take Next-Generation Mobile to the Next Level", *IEEE Vehicular Technology Magazine,* vol. 13, no. 1, pp. 74-84, February 2018.

[20] NGMN Alliance, "Description of Network Slicing Concept", September 2016.

[21] ETSI, "Network Functions Virtualisation (NFV); Virtual Network Functions Architecture", ETSI GS NFV-SWA 001, V1.1.1 (2014-12), European Telecommunications Standards Institute, December 2014.

[22] NGMN alliance, "5G White Paper", 2015.

[23] L. Wang and L. Geng, "Review of Network Slicing Research in Diverse Standards, draft-wang-nfvrg-network-slice-diverse-standards-00, I-D.", July 2018.

[24] 3GPP, "System Architecture for the 5G System", 3GPP TS 23.501, v16.1.0, June 2018.

[25] 3GPP, "Management of 5G networks and network slicing; Concepts, use cases and requirements", 3GPP TS 28.530, v15.1.0, December 2019.

[26] 3GPP, "Management and orchestration of networks and network slicing; Provisioning; Stage 1", 3GPP TS 28.531, v16.2.0, June 2019.

[27] 3GPP, "Management and orchestration of networks and network slicing; Provisioning; Stage 2 and stage 3", 3GPP TS 28.532, v16.0.0, June 2019.

[28] 3GPP, "Study on management and orchestration of network slicing for next generation network", 3GPP TR 28.801, v15.1.0, January 2018.

[29] 3GPP, "Telecommunication management; Architecture", 3GPP TS 32.102, v15.0.0, June 2018.

[30] ITU-T, "Requirements of soft network architecture for mobile", Recommendation ITU-T Y.3323, September 2013.

[31] ITU-T, "High-level technical characteristics of network softwarization for IMT-2020", Recommendation ITU-T Y.3150, September 2017.

[32] ITU-T, "IMT-2020 network management and orchestration framework", Recommendation ITU-T Y.3111, September 2017.

[33] ITU-T, "Framework for the support of multiple network slicing", Recommendation ITU-T Y.3112, May 2018.

© 2018 - 2021 5G-DRIVE Consortium Parties

[34] ETSI, “Network Functions Virtualisation (NFV);Management and Orchestration”, ETSI GS NFV-MAN 001 V1.1.1 (2014-12), European Telecommunications Standards Institute, December 2014.

[35] ETSI, “Network Functions Virtualisation (NFV) Release 3; Evolution and Ecosystem; Report on Network Slicing Support with ETSI NFV Architecture Framework”, ETSI GR NFV-EVE 012 V3.1.1 (2017-12, European Telecommunications Standards Institute, December 2017.

[36] GSM Association, “An Introduction to Network Slicing (white paper)”, 2017.

[37] 5GPPP, “5G PPP Architecture Working Group, View on 5G Architecture (Version 2.0)”, July 18, 2017.

[38] S. Kukliński and L. Tomaszewski, “DASMO: A scalable approach to network slices management and orchestration”, in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, Taipei, 2018.

[39] 3GPP, “Study on Key Quality Indicators (KQIs) for service experience”, 3GPP TR 32.862, v14.0.0, March 2016.

[40] 3GPP, “End-to-end multimedia services performance metrics”, 3GPP TR 26.944, v15.0.0, June 2018.

[41] ITU-T, “Definitions of terms related to quality of service”, ITU-T E.800, September 2018.

[42] GSM Association, “Definition of Quality of Service parameters and their computation”, GSMA IR.42, ver. 9.0, June 2018.

[43] Alliance for Telecommunications Industry Solutions Deployments, “NFV Infrastructure Metrics for Monitoring Virtualized Network”, ATISI0000062, January 2018.

[44] The 5GMonArch Consortium, “Deliverable D6.1 Documentation of Requirements and KPIs and Definition of Suitable Evaluation Criteria”, September 2017.

[45] ETSI, “Network Functions Virtualisation (NFV) Release 3; Management and Orchestration; Requirements and interfaces specification for management of NFV-MANO”, ETSI GS NFV-IFA 031, V3.1.1 (2018-09), European Telecommunications Standards Institute, September 2018.

[46] “OpenStack webpage”, [Online]. Available: https://www.openstack.org. [Accessed March 2019].

[47] Cisco Systems, “High Performance VNFs on NFVI”, [Online]. Available: https://www.cisco.com/c/m/en_us/network-intelligence/service-provider/digital-transformation/high-performance-vnfs-on-nfvi.html . [Accessed March 2019].

[48] Intel Network Builders, “Enhanced Platform Awareness”, [Online]. Available: https://networkbuilders.intel.com/network-technologies/enhancedplatformawareness. [Accessed March 2019].

[49] The 5G ESSENCE Consortium, “D6.2 Prototypes of Network Services and Integration Planning for Use Case 2”, 2019.

[50] “Data Plane Development Kit”, [Online]. Available: http://dpdk.org/ . [Accessed March 2018].

[51] The SHIELD Consortium, “D3.2 Updated specifications, design and architecture of the vNSF ecosystem”, 2018.

[52] Intel Network Builders, “OpenStack EPA feature breakdown and analysis”, [Online]. Available: https://software.intel.com/en-us/articles/openstack-epa-feature-breakdown-and-analysis.

[Accessed March 2019].

[53] ETSI OSM, “Open Source MANO”, [Online]. Available: https://osm.etsi.org/ . [Accessed March 2018].

[54] “Production-Grade Container Orchestration - Kubernetes”, Google Inc., [Online]. Available: https://kubernetes.io/. [Accessed April 2019].

[55] Canonical Ltd., “Juju charms”, [Online]. Available: https://jujucharms.com/. [Accessed March 2019].

[56] “OpenStack Rocky focuses on bare metal clouds, fast forward upgrades and hardware accelerators”, [Online]. Available: https://www.openstack.org/software/rocky/. [Accessed March 2019].

[57] DPDK Documentation, “Enable DPDK in Linux”, [Online]. Available: https://doc.dpdk.org/guides-16.04/linux_gsg/enable_func.html#loading-the-dpdk-kni-kernel-module.

[58] “Compile and Configure OVS & DPDK”, [Online]. Available: https://github.com/openvswitch/ovs/blob/v2.5.0/INSTALL.DPDK.md.

[59] Microsoft Hardware Dev Center, “Single Root I/O Virtualisation (SR-IOV)”, 20 April 2017. [Online]. Available: https://docs.microsoft.com/en-us/windows-hardware/drivers/network/single-root-i-o-virtualization--sr-iov-.

[60] OpenStack Documentation, “Attaching physical PCI devices to guests”, [Online]. Available: https://docs.openstack.org/nova/pike/admin/pci-passthrough.html.

[61] “An advanced guide to OpenStack Networking”, [Online]. Available: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux_openstack_platform/7/html-single/networking_guide/index#enable_the_openstack_networking_sr_iov_agent.

[62] OpenStack Documentation, “PCI Passthrough & SRIOV support”, [Online]. Available: https://wiki.openstack.org/wiki/PCI_passthrough_SRIOV_support.

[63] “Linux Memory Management Documentation: What is NUMA?”, [Online]. Available: https://www.kernel.org/doc/html/v4.18/vm/numa.html#.

[64] “Prometheus webpage”, [Online]. Available: https://prometheus.io. [Accessed March 2019].

[65] “Node Exporter in github”, [Online]. Available: https://github.com/prometheus/node_exporter. [Accessed March 2019].

[66] “OpenStack docs: Welcome to Ceilometer's documentation”, [Online]. Available: https://docs.openstack.org/ceilometer/latest/. [Accessed March 2019].

[67] “OpenStack docs: Shaker”, [Online]. Available: https://docs.openstack.org/performance-docs/latest/test_plans/tenant_networking/shaker.html. [Accessed March 2019].

[68] “Overview - Rally 1.4.1 Documentation”, [Online]. Available: https://rally.readthedocs.io/en/latest/overview/overview.html. [Accessed March 2019].

[69] “Quickstart - os-faults 0.2.1 documentation”, [Online]. Available: https://os-faults.readthedocs.io/en/latest/quickstart/index.html. [Accessed March 2019].

[70] A. Boualouache, S. Senouci and S. Moussaoui, “Towards an Efficient Pseudonym Management and Changing Scheme for Vehicular Ad-Hoc Networks”, in *IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016.

Appendix A OpenStack and OSM installation and configuration

There is extensive documentation on installation and configuration on the OSM & OpenStack Wikis; this information focuses on listing the important configurations and possible pitfalls for 5G-DRIVE.

A. OpenStack installation

OpenStack consists of several independent parts, namely, the OpenStack services. All services authenticate through a common identity service called Keystone. Individual services interact with each other through public APIs, except where privileged administrator commands are necessary. OpenStack services can be composed of several processes. All services have at least one API process, which listens for API requests, pre-processes them and passes them on to other parts of the service. Except for the identity service, the actual work is done by distinct processes. Communication between the processes is handled by message brokers such as AMPQ, RabbitMQ, etc. The minimum requirement for an OpenStack deployment (version Rocky) consists of the following services:

- Identity service – **Keystone**[227]
- Image service – **Glance**[228]
- Compute service – **Nova**[229]
- Networking service – **Neutron**[230]

The OpenStack documentation features separate installation notes for all OpenStack services [24]. Once the minimum requirements have been fulfilled, additional services can be installed. A full list of services exists in the OpenStack documentation. However the ones more relevant to 5G-DRIVE are herein listed:

- Dashboard – **Horizon**[231]
- Block Storage service – **Cinder**[232]
- Object Storage service – **Swift**[233]
- Telemetry Alarming services – **Aodh**[234]
- Telemetry Data Collection service – **Ceilometer**[235]
- Telemetry Event service – **Panko**[236]**.**

The following figure shows the logical architecture of OpenStack services. Mandatory services are considered high-priority. Of the optional ones, medium priority is assigned to the telemetry and visualisation services, while block & object storage is assigned lower priority. Other services are not currently required for 5G-DRIVE, although they can be considered in the future (e.g. Heat/Trove is used for VNF management and orchestration but not utilised in 5G ESSENCE; Sahara[237] is used to

[227] See: *https://docs.openstack.org/keystone/latest/* [Accessed April 2019].
[228] See: *https://docs.openstack.org/glance/latest/* [Accessed April 2019].
[229] See: *https://docs.openstack.org/nova/latest/* [Accessed April 2019].
[230] See: *https://wiki.openstack.org/wiki/Neutron* [Accessed April 2019].
[231] See: *https://docs.openstack.org/horizon/latest/* [Accessed April 2019].
[232] See: *https://wiki.openstack.org/wiki/Cinder* [Accessed April 2019].
[233] See: *https://docs.openstack.org/ocata/cli-reference/swift.html* [Accessed April 2019].
[234] See: *https://docs.openstack.org/ocata/cli-reference/aodh.html* [Accessed April 2019].
[235] See: *https://docs.openstack.org/ceilometer/pike/install/get_started.html* [Accessed April 2019].
[236] See: *https://docs.openstack.org/panko/latest/* [Accessed April 2019].
[237] See: *https://wiki.openstack.org/wiki/Sahara*

provision data processing frameworks like *Apache Hadoop, Spark, Storm*, etc.; Ironic[238] is used for bare metal provisioning, etc.).

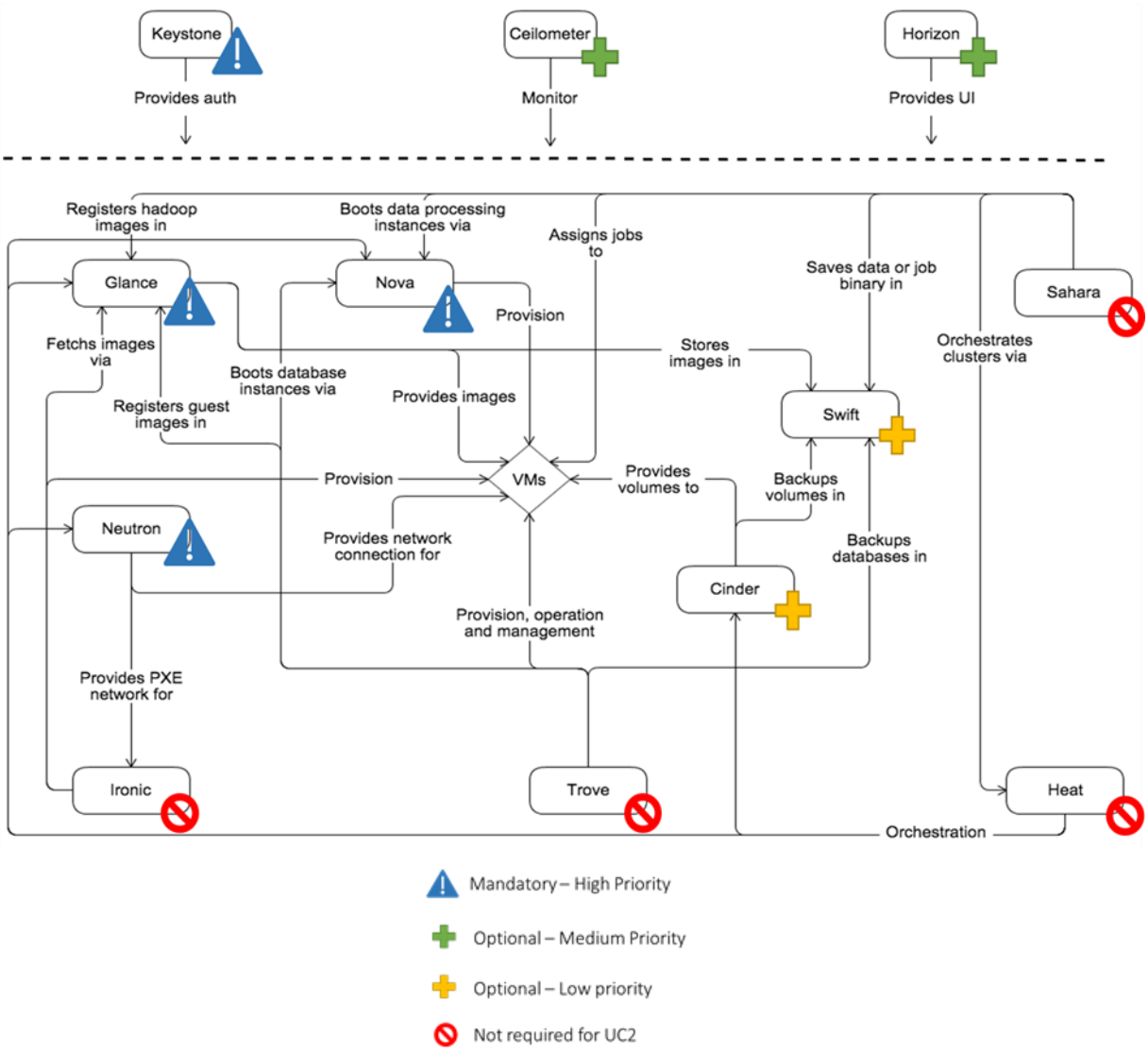


Figure 60: OpenStack logical architecture, showing prioritisation of services for the 5G-DRIVE.

B. OSM installation

OSM can be installed in a single VM, with the following requirements [6]:

- MINIMUM: 2 CPUs, 4 GB RAM, 20GB disk and a single interface with Internet access.
- RECOMMENDED: 2 CPUs, 8 GB RAM, 40GB disk and a single interface with Internet access.
- Ubuntu16.04 [6] (64-bit variant required) as base image.

[238] For further details also see: *https://wiki.openstack.org/wiki/Ironic*

Since OSM creates a dedicated container to house the Juju proxy charm (required for the configuration and lifecycle management of each VNF instance), it is necessary to select a larger flavour to provide storage for multiple containers. Larger storage is also essential for future upgrades to the OSM version.

Hence, any proper 5G-DRIVE deployment will rely on the recommended settings, and a minimum flavour is considered inadequate. Additional features can be added during installation, as shown in the following table.

Table 16: Additional OSM components that can be enabled during installation.

Additional component	Description	Priority
Vim-emu	Execution of network functions packaged as Docker containers, in emulated network topologies running locally on the developer's machine.	Low – UC2 does not require Docker-based microservices
Performance Management Stack	Metrics collection	High – Integration of VIM & VNF metrics & visualisation
ELK	Fault Management & Events Visualisation	Medium – Useful functionality but still in the experimental phase
Juju metrics	Information gathering from the VNFs	Low – Monitoring of critical VNFs is already gathered in *Prometheus*, although this can be added for other VNFs that do not already feature node_exporter.

C. OpenStack configuration

It is important to note that the VIM needs to be configured properly in order to be reachable from OSM and to support some required features, following a specified procedure. First, the admin needs to guarantee that OpenStack API endpoints are reachable from OSM and specifically from the Resource Orchestration (RO) container. Then, it is necessary to create a management network, with DHCP enabled, reachable from OSM. The network is used by the VCA (Juju) for configuring the VNFs once they are running. It is recommended to create a provider network, isolated from OpenStack. The next step is to create a valid tenant/user with rights to create/delete flavours. The easiest way is to create a user and assign it the role “admin”. Another option is to change the general flavour management policies by configuring Nova to allow flavour creation per user. Finally, the VNF images need to be uploaded to the VIM storage. Another important aspect is that OpenStack and OSM support Enhanced Platform Awareness (EPA). EPA is a contribution to OpenStack by Intel. EPA aims to expose hardware characteristics of the NFVI to the VIM, enabling hardware features that can be used to increase the performance of VNFs (e.g. SR-IOV & DPDK) and make them visible to OSM for fine-grained deployment. Each selected EPA feature will thus need to be configured on the NFVI and the VIM.

D. OSM installation testing

After the installation has completed, it is necessary to test that the important OSM services are running. After release FIVE, most services are dockerised. As a result of the installation, thirteen containers are created in the host.

OSM deploys a Juju controller in a dedicated LXD container. Known installation issues include misconfiguration of LXD and Juju, in which case, both LXD and the Juju controller need to be re-installed. Furthermore, it is important to secure the OSM installation by changing the default user account credentials (user: admin, password: admin).

E. Adding a VIM account to OSM and Enabling VIM metrics

Assuming that OpenStack has been installed and properly configured to be reachable by OSM, it can be added to OSM either via the OSM client's command line interface or via the web interface. It is important to note that if an SDN controller is added in OSM, it needs to also be associated with the VIM.

If OpenStack telemetry is installed (i.e.: Ceilometer, Aodh, Gnocchi), metric collection can be configured on the OSM MON (monitoring) container and implemented through a dedicated Kafka bus[239]. As of Release 5.0.0, the metric collection has been tested with OpenStack VIM with Keystone v3 authentication[240] and legacy or Gnocchi-based telemetry services[241]. Before activating metrics, it is important to make the MON container aware of the granularity of the Telemetry system (e.g. sample rate).

F. Onboarding the VNF

The VNF image must be first uploaded to the VIM. In OpenStack, this is performed with the OpenStack image create command. Then the administrator must upload the VNFD and NSD to OSM. This can easily be performed via the web interface or through the CLI of the OSM Client. Once the related descriptors are uploaded, the VNF is ready for instantiation as long as there are resources available in the NFVI-PoP. A common pitfall also lies with the configuration of OSM during installation: since a dedicated lxd container is created for each instance to host the VNFM, there need to be enough disk space resources to support multiple concurrent instantiations. In any other case, the instantiation will fail until more space is freed within OSM. It is also worth noting, that there are conversion tools available to port descriptors from older versions of OSM to release FIVE, thus facilitating the use of existing VNFs.

[End of document]

[239] See, *for example*: *https://kafka.apache.org/*
[240] See: *https://docs.openstack.org/python-keystoneclient/latest/api/keystoneclient.auth.identity.v3.html*
[241] See: *https://cloud.epam.com/site/competency_center/e=p=c_services/telemetry_as_a_service_(=t=m=s)*