ETSI GS NIN 006 V1.1.1 (2024-07)

Non-IP Networking (NIN);
Scenario Definitions of Next Generation Protocols (NGP)

Disclaimer

The present document has been produced and approved by the Non-IP Networking (NIN) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGS/NIN-06v111

Keywords
core network, cyber security, IoT, mobility, network, QoE, reliability, security, service, use case

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights 7
Foreword 7
Modal verbs terminology 7
1 Scope 8
2 References 8
2.1 Normative references 8
2.2 Informative references 9
3 Definition of terms, symbols and abbreviations 11
3.1 Terms 11
3.2 Symbols 18
3.3 Abbreviations 18
4 Overview 21
5 Issues to be addressed by the Scenarios 23
6 Model References 25
6.0 Introduction 25
6.1 LTE Mobile Network Model 25
6.2 L2 and L3 VPN services 27
6.2.0 Introduction 27
6.2.1 MPLS/BGP Layer 3 Virtual Private Networks 28
6.2.2 VPLS, Virtual Private Line Services and Ethernet-VPN 29
6.3 All IP Core Network Model 31
6.4 NFV Reference Model 33
6.5 MEC Reference Model 34
7 Referenced Use Cases 34
8 Scenarios 35
8.1 Addressing 35
8.1.0 Introduction 35
8.1.1 Model Architecture 36
8.1.2 Scenario Description 37
8.1.2.0 Introduction 37
8.1.2.1 Scenarios for mobile communication 37
8.1.2.2 Scenarios for multi-homing and load balancing 38
8.1.3 Applicable Issues 39
8.1.4 Applicable Use Cases 39
8.1.4.1 Case 1: UE communicates with a fixed device; UE is moving within a same P-GW domain 39
8.1.4.2 Case 2: UE communicates with a fixed device; UE is moving across different P-GW domain 40
8.1.4.3 Case 3: UE communicates with a fixed device; UE is moving across heterogeneous access network 40
8.1.4.4 Case 4: UE communicates with another UE; UE is moving within a same P-GW domain 41
8.1.4.5 Case 5: UE communicates with another UE; UE is moving across different P-GW domain 41
8.1.4.6 Case 6: UE communicates with another UE; UE is moving across heterogeneous access network 42
8.1.4.7 Case 7: Multi-homing host connected to different ISP for link protection or load balance 42
8.1.4.8 Case 8: Customer network with multi-homing site connected to different ISP for link protection or load balancing 42
8.1.5 Scenario Targets 43
8.2 Security 44
8.2.1 Model Architecture/Protocol Stacks 44
8.2.2 Scenario Description 44
8.2.2.1 Scenario summary 44
8.2.2.2 Security approach 44
8.2.2.3 Description of new security challenges 45
8.2.3 Applicable Issues 45

8.2.4 Applicable Use Cases 49
8.2.5 Scenario Targets 49
8.3 Mobility 51
8.3.1 Model Architecture 51
8.3.2 Scenario Description 53
8.3.3 Applicable Issues 54
8.3.4 Applicable Use Cases 56
8.3.4.0 Introduction 56
8.3.4.1 Case 1: Multi-Access, Session & Bearer connection, Same Macro 56
8.3.4.2 Case 2: Multi-Access, Session & Bearer connection, with Macro HO 56
8.3.4.3 Case 3: Single Access, Session & Bearer, Same Macro 56
8.3.4.4 Case 4: Single Access, Multi-Session, Multi-Bearer, Same Macro 56
8.3.4.5 Case 5: Fast, Single Access, Multi-Session, Multi-Bearer, with Macro HO 57
8.3.4.6 Case 6: Fast, Multi-Access, Session & Bearer connection, with Macro HO 57
8.3.4.7 Case 7: Fast, Multi-Access, Session & Bearer connection, with Macro HO 57
8.3.5 Scenario Targets 57
8.4 Multi-Access Support (including FMC) 58
8.4.1 Model Architecture 58
8.4.2 Scenarios 58
8.4.3 Scenario Description 59
8.4.4 Applicable Issues 60
8.5 Context Awareness 60
8.5.1 Model Architecture/Protocol Stacks 60
8.5.2 Scenario Description 61
8.5.3 Applicable Issues 64
8.5.4 Applicable Use Cases (from Annex A) 65
8.5.5 Scenario Targets 65
8.6 Performance Improvement & Content Enablement 66
8.6.1 Model Architecture 66
8.6.2 Scenario Descriptions 68
8.6.2.0 Introduction 68
8.6.2.1 Scenario #1 - Adaptive video streaming 69
8.6.2.2 Scenario #2 - 8K Video Streaming 69
8.6.2.3 Scenario #3 - Live Virtual Reality 70
8.6.2.4 Scenario #4 - URLLC For Time-Critical IoT 70
8.6.3 Issues with TCP Congestion Control 70
8.6.3.1 An appraisal of Congestion Management 70
8.6.3.2 An Introduction to Current TCP Congestion Mechanisms 71
8.6.4 Applicable Issues and Recommendations 72
8.6.5 Applicable Use Cases (from Annex A) 74
8.6.5.0 Introduction 74
8.6.5.1 Case 1: New Transport Protocol 74
8.6.5.2 Case 2: Use Case for Flexible Application Traffic Routing 74
8.6.5.3 Case 3: In-Network Caching 74
8.6.5.4 Case 4: Deterministic Network Reporting/Profiling 74
8.6.6 Scenario Targets 74
8.7 Network Virtualisation 75
8.7.0 Introduction 75
8.7.1 Model Architecture 76
8.7.2 Scenario Description 80
8.7.2.1 Scenario #1: Network Virtualisation in EPS 80
8.7.2.2 Scenario #2: Virtualised RAN 81
8.7.3 Applicable Issues 82
8.7.4 Applicable Use Cases 84
8.7.4.1 Case 1: Network Slicing 84
8.7.4.2 Case 2: Network Slicing: With Simultaneous access to different instances of Virtualised core 86
8.7.4.3 Case 3: MEC and Network Virtualisation 86
8.7.4.4 Case 4: Cloud interconnect (Mobile/Fixed networks) 86
8.7.4.5 Case 5: C-RAN Enhanced Computational Flexibility 87
8.7.4.6 Case 6: Heterogeneity of RAT 88
8.7.4.7 Case 7: Performance Enhancement of Low-power RRU 88
8.7.5 Scenario Targets 88

8.8 IoT Scenario 90
8.8.1 Model Architecture/Protocol Stacks 90
8.8.2 Scenario Descriptions 90
8.8.2.0 Introduction 90
8.8.2.1 Active Assisted Living (AAL) 91
8.8.2.2 Cooperation between factories and remote applications 91
8.8.2.3 Smart glasses in industrial applications 91
8.8.3 Applicable Issues 91
8.8.4 Applicable Use Cases (from Annex A) 93
8.9 Energy Efficiency 93
8.10 eCommerce 94
8.11 Mobile Edge Computing (MEC) 94
8.11.0 Introduction 94
8.11.1 Model Architecture 95
8.11.2 Applicable Issues and Recommendations 96
8.11.3 Applicable Use Cases 97
8.11.3.0 Introduction 97
8.11.3.1 Case 1: Video Stream Analysis service 97
8.11.3.2 Case 2: Augmented and Virtual Reality service 98
8.11.3.3 Case 3: Assistance for intensive computation service 98
8.11.3.4 Case 4: IoT Gateway service 98
8.11.3.5 Use Case 5: Connected Vehicles service scenario 99
8.11.4 Scenario Targets 100
8.12 Mission Critical Services: PSC and PUC 100
8.12.0 Introduction 100
8.13 Drones, Autonomous and Connected Vehicles 100
8.13.0 Introduction 100
8.13.1 Model 100
8.13.2 Scenarios 101
8.13.3 Applicable Issues and Recommendations 102
8.13.4 Applicable Use Cases 104
8.13.4.0 Introduction 104
8.13.4.1 Hazardous operations 104
8.13.4.2 Driverless vehicles 105
8.13.4.3 Automated Convoys ('platooning') 105
8.13.4.4 Connected vehicles 105
8.14 URLLC: Ultra-Reliable and Low Latency Communications 105
8.14.0 Introduction 105
8.14.1 Model architecture 106
8.14.2 Scenario description 107
8.14.2.0 Scenarios Introduction 107
8.14.2.1 Handover interruption caused by UE's mobility 107
8.14.2.2 Interruption Latency caused by mobility of application 108
8.14.2.3 E2E latency enlarged by Indeterminate processing delay in network nodes 108
8.14.2.4 Conflict between high reliability and low latency under wireless packet loss 109
8.14.3 Applicable Issues and Recommendations 109
8.14.4 Use cases 109
8.13.4.1 Case 1: Local UAV Collaboration 109
8.14.4.2 Case 2: Industrial Factory Automation 110
8.14.4.3 Case 3: V2X 110
8.14.4.4 Scenario Targets 110

Annex A (informative): Use Cases & Parameterization 111

Annex B (informative): 5G mobile network model 126

B.0 Introduction 126
B.1 References 126
B.2 Reading the scenarios in the context of 5G 126
B.3 Key differences from LTE mobile Network Model 126

B.4 Service Based Architecture 126
B.5 5G Protocol stacks 127
B.5.0 Introduction 127
B.5.1 User plane protocols 127
B.5.2 5G Control Plane Protocols 128
B.6 Networking issues carried over from LTE architecture 129
Annex C (informative): Void 130
History 131

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Group Specification (GS) has been produced by ETSI Industry Specification Group (ISG) Non-IP Networking (NIN).

NOTE: This Group Specification (GS) has been initially produced by ETSI Industry Specification Group (ISG) Next Generation Protocols (NGP) as ETSI GS NGP 001, and revised by ETSI Industry Specification Group (ISG) Non-IP Protocols (NIN) .

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The scope of the present document is to specify the minimum set of key scenarios for the Next Generation Protocols (NGP), Industry Specific Group (ISG).

2 References

2.1 Normative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

Referenced documents which are not found to be publicly available in the expected location might be found at https://docbox.etsi.org/Reference/.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are necessary for the application of the present document.

[1] NGMN: "5G Whitepaper".

[2] Recommendation ITU-T Y.2091: "Terms and definitions for next generation networks".

[3] Recommendation ITU-T Y.2720: "NGN identity management framework".

[4] IETF RFC 8113: "Locator/ID Separation Protocol (LISP): Shared Extension Message & IANA Registry for Packet Type Allocations".

[5] IETF RFC 760: "DoD standard Internet Protocol".

[6] ISO/IEC 7498-1:1994: "Information technology - Open Systems Interconnection -- Basic Reference Model: The Basic Model".

[7] Department of Defense World Geodetic System 1984 TR 8350.2.

[8] ETSI GS NFV 002: "Network Functions Virtualisation (NFV); Architectural Framework".

[9] ETSI GS NFV 003: "Network Functions Virtualisation (NFV); Terminology for Main Concepts in NFV".

[10] IETF RFC 4364: "BGP/MPLS IP Virtual Private Networks (VPNs)".

[11] IETF RFC 4761: "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling".

[12] IETF RFC 3753: "Mobility Related Terminology".

[13] IETF RFC 7333: "Requirements for Distributed Mobility Management".

[14] IETF draft-ietf-lisp-lcaf-14 (LISP): "LISP Canonical Address Format (LCAF)".

[15] IETF draft-farinacci-lisp-eid-anonymity-00 (LISP): "LISP EID Anonymity".

[16] ETSI GS NFV 001 (V1.1.1): "Network Functions Virtualisation (NFV); Use Cases".

[17] Void.

[18] ETSI GS NFV-SEC 003 (V1.1.1): "Network Functions Virtualisation (NFV); NFV Security; Security and Trust Guidance".

[19] ETSI GS MEC 001 (V1.1.1): "Mobile Edge Computing (MEC) Terminology".

[20] ETSI GS MEC 003 (V1.1.1): "Mobile Edge Computing (MEC); Framework and Reference Architecture".

[21] ETSI GS MEC-IEG 004 (V1.1.1): "Mobile-Edge Computing (MEC); Service Scenarios".

[22] ETSI TS 103 307: "CYBER; Security Aspects for LI and RD Interfaces".

[23] ETSI GS NFV-SEC 009 (V1.1.1): "Network Functions Virtualisation (NFV); NFV Security; Report on use cases and technical approaches for multi-layer host administration".

[24] ETSI TS 132 500: "Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements (3GPP TS 32.500)".

[25] MEC White-paper: "Mobile Edge Computing: A key technology towards 5G", 2015.

[26] IEEE 802.1Q™-2011: " IEEE Standard for Local and metropolitan area networks--Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks".

[27] ETSI TS 123 401: "LTE; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (3GPP TS 23.401 Release 13)".

[28] ETSI TS 122 261: "5G; Service requirements for next generation new services and markets (3GPP TS 22.261 Release 15)".

[29] ETSI TS 122 280: "LTE; Mission Critical Services Common Requirements (3GPP TS 22.280 Release 14)".

[30] Society of Automotive Engineers, J3016: "Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems".

[31] IETF RFC 3246: "An Expedited Forwarding PHB (Per-Hop Behavior)".

[32] ETSI TS 123 501: "5G; System Architecture for the 5G System (3GPP TS 23.501 Release 15)".

[33] 3GPP TR 29.891: "3rd Generation Partnership Project; Technical Specification Group Core Network and Terminals; 5G System - Phase 1; CT WG4 Aspects (Release 15)".

[34] ETSI TS 138 300: " 5G; NR; Overall description; Stage-2 (3GPP TS 38.300)".

[35] ETSI TS 129 281: "Universal Mobile Telecommunications System (UMTS); LTE; General Packet Radio System (GPRS) Tunnelling Protocol User Plane (GTPv1-U) (3GPP TS 29.281)".

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] 3GPP TR 22.891: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers; Stage 1".

[i.2] 3GPP TR 23.799: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on Architecture for Next Generation System (NexGen)".

[i.3] ETSI TR 121 905: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Vocabulary for 3GPP Specifications (3GPP TR 21.905)".

[i.4] 5GPPP Whitepaper on Automotive Vertical Sector.

[i.5] 5GPPP Whitepaper on Energy Vertical Sector.

[i.6] 5GPPP Whitepaper on Factories of the Future.

[i.7] 5GPPP Whitepaper on E-Health.

[i.8] Elements of Mathematics: "General Topology", Berlin, Springer- Verlag, 1990, Bourbaki, N. 1971.

[i.9] "Elements of the Topology of Plane Sets of Points", Newman, M, 1964.

[i.10] Stallings, William: "High-Speed Networks and Internets", Prentice-Hall™, 2002.

[i.11] Risk Nexus: "Overcome by cyber risks? Economic benefits and costs of alternate cyber futures".

[i.12] "A Binary Feedback Scheme for Congestion Avoidance in Computer Networks with Connectionless Network Layer", ACM Transactions on Computer Systems, Vol. 8, No. 2, May 1990, pp. 158-181, K. Ramakrishnan and Raj Jain.

[i.13] Digital Equipment Corporation Technical Report No. DEC-TR-510: "Congestion Avoidance in Computer Networks with A Connectionless Network Layer: Part IV: A Selective Binary Feedback Scheme for General Topologies", August 1987, 43 pp., K. Ramakrishnan and Raj Jain.

[i.14] Void.

[i.15] IETF RFC 4762: "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling".

[i.16] IETF RFC 4984: "Report from the IAB Workshop on Routing and Addressing".

[i.17] 3GPP TR 23.863: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Support of Short Message Service (SMS) in IP Multimedia Subsystem (IMS) without Mobile Station International ISDN Number (MSISDN); Stage 2".

[i.18] 3GPP TR 22.864: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers - Network Operation; Stage 1".

[i.19] IETF RFC 6582: "The NewReno Modification to TCP's Fast Recovery Algorithm".

[i.20] IETF RFC 2018: "TCP Selective Acknowledgment Options".

[i.21] ETSI GS MEC 002: "Mobile Edge Computing (MEC); Technical Requirements".

[i.22] ETSI GS MEC-IEG 005: "Mobile-Edge Computing (MEC); Proof of Concept Framework".

[i.23] IETF RFC 7041: "Extensions to the Virtual Private LAN Service (VPLS) Provider Edge (PE) Model for Provider Backbone Bridging".

[i.24] 5G Manifesto for timely deployment of 5G.

[i.25] ETSI TR 138 913: "5G; Study on Scenarios and Requirements for Next Generation Access Technologies (3GPP TR 38.913)".

[i.26] IETF Charter of IETF DMM documents.

[i.27] Broadband Forum TR-069: "CPE WAN Management Protocol".

[i.28] 3GPP TR 38.801: "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on new radio access technology: Radio access architecture and interfaces".

[i.29] 3GPP TR 36.881: "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Study on latency reduction techniques for LTE".

[i.30] Ericsson Research AB: "Service Mobility in Mobile Networks", 2015 IEEE 8th International Conference on Cloud Computing .

[i.31] Nokia: "W01-Third Workshop on 5G Architecture (5GArch 2016)", Mobility Management Enhancements for 5G Low Latency Services.

[i.32] IETF draft-briscoe-tsvwg-ecn-l4s-id-00 K. De Schepper, I. Tsang, Bell Labs, B. Briscoe, Ed Simula Research Lab: "Identifying Modified Explicit Congestion Notification (ECN) Semantics for Ultra-Low Queuing Delay".

[i.33] Void.

[i.34] 3GPP TR 22.862: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Feasibility Study on New Services and Markets Technology Enablers for Critical Communications; Stage 1" (Release 14).

[i.35] National Highway Traffic Safety Administration: "Preliminary statement of policy concerning automated vehicles" (2013).

[i.36] 5GPPP white paper: "5G Automotive Vision".

[i.37] 3GPP TR 22.886: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on enhancement of 3GPP Support for 5G V2X Services (Release 15)".

[i.38] Society of Automotive Engineers Mobulus: "Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems".

[i.39] Driverless Future.

[i.40] ETSI GS NFV 006 (V4.4.1) : "Network Functions Virtualisation (NFV) Release 4; Management and Orchestration; Architectural Framework Specification".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the terms applying to scenarios that include mobile network architectures given in ETSI TR 121 905 [i.3] and 3GPP TR 23.799 [i.2] and the following apply:

access point: point of access to a network, which in this generic NGP context may be a traditional Wi-Fi access point, 3GPP cellular network base station, RRU supporting a cell or sector or part thereof if the cell is configured as a multi-point access cell

address: identifier for a specific termination point and is used for routing to this termination point

NOTE: See Recommendation ITU-T Y.2091 [2].

application process: instantiation of a program executing in a processing system intended to accomplish some purpose

NOTE: An application contains one or more application protocol machines.

application process name: name of an application process

application protocol: protocol characterized by modifying state external to the protocol by performing remote operations on an object model

NOTE: The minimal set of operations are create/delete, start/stop and read/write.

application protocol name: name of an application protocol

asymmetric link: link with transmission characteristics which are different depending upon the relative position or design characteristics of the transmitter and the receiver of data on the link

NOTE: For instance, the range of one transmitter may be much higher than the range of another transmitter on the same medium see IETF RFC 3753 [12].

autonomous: entity capable of piloting itself based on sensory input and pre-defined behaviours, including collision avoidance, speed limits and geographical constraints

NOTE 1: Used in the context of communications pertaining to an autonomous drone or vehicle in the present document.

NOTE 2: However, a remote piloting capability may be provided via a network to which the vehicle is able to communicate with.

autonomous drone: autonomous vehicle with no human operator on-board

NOTE: The distinction between a 'drone' and an 'autonomous drone' is that in the case of an 'autonomous drone', the vehicle is piloted through on-board sensor processing, and optionally, and less frequently remote control 'managed by' a human operator. Where: the term 'managed by' indicates that the human operator may be actively monitoring the vehicle's progress, and taking control manually on an event triggered basis, as necessary; or delegating the remote control to a computing process.

autonomous vehicle: vehicle capable of piloting itself, that also has a human operator on-board

NOTE 1: The distinction between an 'autonomous vehicle' and an 'autonomous drone' is the presence of a human operator in the vehicle.

NOTE 2: The 'autonomous vehicle' is piloted through a combination of on-vehicle sensor readings, which are processed to determine action, and optional interjection from a human operator. An example of this kind of situation is when there is a 'self-driving car' with a human passenger on-board who is capable of piloting the car.

backhaul: transmission system between a base station entity and the cellular core network or Non-Access Stratum

binding a name to an object: function, $Fn(M_{NS})$, that defines the mapping of elements of NS(namespace) to elements of M(object)

NOTE 1: The result of this function is called a *binding*. e.g. In LISP, the binding operation is called mapping.

NOTE 2: For example, <ID1, RLOC1> is the mapping of ID1="identity1" to RLOC1="an ip address or any other form of addressing".

care-of-address: IP address associated with a mobile node while visiting a foreign link; the subnet prefix of this IP address is a foreign subnet prefix

NOTE: A packet addressed to the mobile node which arrives at the mobile node's home network when the mobile node is away from home and has registered a Care-of Address will be forwarded to that address by the Home Agent in the home network see IETF RFC 3753 [12].

centralized mobility management: makes use of centrally deployed mobility anchors

NOTE: Please see IETF RFC 7333 [13].

compound connection: connection that includes logical connectivity to more than one access network at a time

congestion avoidance: mechanism that operates the network at the knee of the congestion or response time (or delay) curve to optimize the trade-off between response time and throughput

congestion 'cliff': congestion point of the response time (or delay) curve at which a session collapses

congestion control: control scheme that manages packet congestion, by constantly testing the congestion level that causes the network communications to collapse and responds by introducing packet loss to reduce the load during periods of congestion so that the network can recover to an uncongested state

NOTE 1: The congestion level that causes the network to collapse is often referred to as the 'congestion cliff' because the curve of throughput versus load, becomes very steep beyond this congestion level.

NOTE 2: Addresses the "social" problem of having various logical links in a network cooperate in order to avoid and/ or recover from congestion of the intermediate nodes that they share.

congestion 'knee': congestion point of the response time (or delay) curve at which as session begins to notably deteriorate

connected vehicle: vehicle connected to one or more communications networks and piloted by a human operator

NOTE 1: The network will allow the vehicle to share data with other connected vehicles and remote servers.

NOTE 2: A Connected vehicle does not need to be autonomous, but is network connected for purposes of assisted navigation, environmental updates, vehicle analysis, infotainment etc.

connection: shared state between EFCPM-instances

NOTE: See ISO/IEC 7498-1 [6].

C-RAN: Cloud RAN where the physical radio part of a base station termed the RRU has been remoted from its base band equipment termed the BBU via 'fronthaul' transmission and the BBU part connects the composite RAN equipment to the cellular core via 'backhaul'

NOTE: Often multiple RRU communicate with a single BBU to effect RAN optimization at the BBU level across a number of Cells provided by the RRH.

data transfer protocol, machine dtp(m): half of the EFCP that performs tightly bound mechanisms, such as ordering, and fragmentation/reassembly

NOTE: One instantiation is created for each flow allocated, see ISO/IEC 7498-1 [6].

Data Transfer Control Protocol, Machine DTCP(M): half of the EFCP that performs loosely bound (feedback) mechanisms, such as retransmission and flow control

NOTE: This protocol maintains state, which is discarded after long periods of no traffic (2MPL). One instantiation is created for each flow requiring either flow control or retransmission control. See ISO/IEC 7498-1 [6].

distance vector: characteristic of some routing protocols in which, for each desired destination, a node maintains information about the distance to that destination, and a vector (next hop) towards that destination

NOTE: See IETF RFC 3753 [12].

distributed application: collection of cooperating APs that exchange information using IPC and maintain shared state

distributed mobility management: not centralized, so that traffic does not need to traverse centrally deployed mobility anchors far from the optimal route

NOTE: See IETF RFC 7333 [13].

drone: powered vehicle remotely controlled by a human operator

NOTE: To distinguish these entities from radio-controlled planes or cars used by hobbyists, the drone should be a client and/or server of data towards a network: for example, it may send video or sensor readings, or be sent images of objects it will aim to detect in its environment.

D-RAN: traditional RAN where the physical radio part of a base station and its base band equipment are co-located at the base station cell site and connected to the rest of the cellular network with 'backhaul' transmission

dual connectivity: mechanism whereby a device can access multiple cells/access points at the same time to bond multiple single cell/access point capabilities together to increase available throughput

Endpoint ID (EID): in LISP binding operation and is called a mapping

NOTE: For example <ID1, RLOC1> is the mapping of ID1="identity1" to RLOC1="an IP address or any other form of addressing", see IETF RFC 6830 [4], [14] and [15].

Error and Flow Control Protocol (EFCP): data transfer protocol required to maintain an instance of IPC within a layer characterized by modifying state internal to the protocol

NOTE: The functions of this protocol may ensure reliability, order, and flow control as required.

Error and Flow Control Protocol Machine (EFCPM): task that instantiates an instance of the EFCP for a single flow or connection

NOTE: An EFCPM consists of two state machines loosely coupled through a single state vector: one that performs the tightly bound mechanisms, referred to as the Data Transfer PM; and the other that performs the loosely coupled mechanisms, referred to as the Data Transfer Control PM, see ISO/IEC 7498-1 [6].

flooding: process of delivering data or control messages to every node within the network under consideration

NOTE: See IETF RFC 3753 [12].

flow control: often referred to as ETE Flow control, see definition in ETSI TR 121 905 [i.3]

front-haul: transmission between separated component parts of a traditional base station when it has been functionally split into at least 2 parts and those parts are remote from each other

function chaining: virtual inter-connection of VNFs to form a NS

graph: ordered pair G = (V, E) comprising a set V of vertices or nodes or points together with a set E of edges or arcs or lines, which are 2-element subsets of V

NOTE: I.e. an edge is related with two vertices, and the relation is represented as an unordered pair of the vertices with respect to the particular edge).

grouping service slice: service chain built to render support for a virtual service offering according to a defined subscriber grouping

NOTE: See 3GPP TR 23.799 [i.2] for further information on the 3GPP ongoing definition of Network Slicing.

handover: process by which an active Mobile Node (in the Active State) changes its point of attachment to the network, or when such a change is attempted

NOTE: The access network may provide features to minimize the interruption to sessions in progress. This procedure is also called hand-off. See IETF RFC 3753 [12].

home address: IP address assigned to a mobile node, used as the permanent address of the mobile node

NOTE: This address is within the mobile node's home link. Standard IP routing mechanisms will deliver packets destined for a mobile node's home address to its home link. See IETF RFC 3753 [12].

Hybrid RAN (H-RAN): optimized form of RAN using concepts from both C-RAN and D-RAN

identifier: series of digits, characters and symbols or any other form of data used to identify subscriber(s), user(s), network element(s), function(s), network entity(ies) providing services/applications, or other entities (e.g. physical or logical objects)

NOTE: See Recommendation ITU-T Y.2720 [3].

identity: information about an entity that is sufficient to identify that entity in a particular context

NOTE: See Recommendation ITU-T Y.2720 [3].

Instance Identifier (ID): used to define extended forms of EID as a multi-tuple value

NOTE: Where (IID, EID) is one example of an extended EID, see IETF RFC 6830 [4].

IoT(mobileS): mobile capable IoT device with one or more sensors

IoT(mobileSA): mobile capable IoT device with one or more sensors and one or more actuators

IoT(staticS): static capable IoT device with one or more sensors

IoT(staticS): static capable IoT device with one or more sensors and one or more actuators

(IP) address: shorthand for Internet Protocol address

NOTE: See IETF RFC 760 [5].

IPC-process: application process that is a member of (N)-layer and implements locally the functionality to support IPC using multiple subtasks

NOTE: Specific for a layer (N).

link: communication facility or physical medium that can sustain data communications between multiple network nodes, such as an Ethernet simple or bridged)

NOTE: A link is the layer immediately below IP. In a layered network stack model, the Link Layer (Layer 2) is normally below the Network (IP) Layer (Layer 3), and above the Physical Layer (Layer 1), see IETF RFC 3753 [12].

local broadcast: delivery of data to every node within range of the transmitter

NOTE: See IETF RFC 3753 [12].

(2D Geographic) Location: specifies the physical location of a 2D point on the earth using two coordinates:

i) latitude and

ii) longitude

NOTE: As referenced in World Geodetic System 1984 [7].

(3D Geographic) Location: 2D location specified with an accompanying altitude expressed as metres above sea level or (ASL)

mobility management: solutions that lie at the centre of the wireless Internet and enable mobile devices to partake in IP networks anytime and anywhere

NOTE: See IETF Charter of IETF DMM WG [i.26]. Includes the setup, maintenance(handover) and release of various physical radio resources when the mobility management is operated with at least one end of a group of communicating peers are attached to the network via an air interface.

(N)-address: identifier that is a synonym for the IPC-Process-Instance, which is a member of a (N)-layer

NOTE 1: An address is only unambiguous within the (N)-layer (and assigned by the (N)-layer).

NOTE 2: This identifier may be assigned to facilitate the operation of the (N)-layer, i.e. location-dependence for routing, see ISO/IEC 7498-1 [6].

(N)-API-primitive: library or system call for a (N)-layer used by an application-process to invoke system functions, in particular IPC functions, such as requesting the allocation of IPC resources

NOTE: See ISO/IEC 7498-1 [6].

N-Concurrent Multi-Access Network: eco-system that includes more than one access network and allows a user device to connect concurrently with more than one of these networks at a time

(N)-Connection-endpoint-id: identifier unambiguous within the scope of an IPC Process that identifies an EFCPM-instance

NOTE: In the Internet, port-id = CEP-id = socket. This also creates several security vulnerabilities, see ISO/IEC 7498-1 [6].

(N)-connection-identifier: identifiers internal to the (N)-layer and unambiguous within the scope of two communicating EFCPMs of that layer

NOTE: The (N)-connection-identifier is commonly formed by the concatenation of the source and destination CEP-ids to identify the two directions of the connection, see ISO/IEC 7498-1 [6].

(N)-Flow: binding of source and destination (N)-connection-endpoints to source and destination (N)-ports

(N)-Layer: collection of processes cooperating as a distributed application to provide inter-process communication (IPC), that create a locus of distributed shared state of a given scope

NOTE: Layers of different rank will generally have different scope.

name: unique string, N, in some alphabet, A, that unambiguously denotes some object or denotes a statement in some language, L

NOTE 1: The statements in L are constructed using the alphabet, A.

NOTE 2: May be mapped to an address by a process or application, see Recommendation ITU-T Y.2091 [2].

name-space: set {N} of names from which all names for a given collection of objects are taken

NOTE 1: A function, M_{NS}, which defines the class of objects, M, that may be named with elements of NS.

NOTE 2: This is referred to as the scope of the name space. This may refer to actual objects or the potential for objects to be created. A name from a given name space may be bound to one and only one object at a time.

neighbour/neighbor: any other node to which data may be propagated directly over the communications medium without relying on the assistance of any other forwarding node

NOTE: As defined in IETF RFC 3753 [12].

neighbourhood/neighborhood: all the nodes which can receive data on the same link from one node whenever it transmits data

NOTE: As defined in IETF RFC 3753 [12].

(Network) Graph: graph of a network

NOTE: A mathematical description of a network by means of two entities:

i) vertices, which represents the nodes; and

ii) edges, which represents their interconnections (see mathematical definition of Graph above).

network slice: function chain built to render a virtual network grouping to support a defined Subscriber

network slicing (dedicated): network slicing capability where the resources of the network slice are allocated for the duration of the session

EXAMPLE: A VR session.

network slicing (dynamic priority): network slicing capability where the resources of the network slice are allocated so as to be able to be reserved for given period of time, according to an indicated priority

EXAMPLE: For a particular live broadcast.

network slicing (reserved): network slicing capability where the resources of the network slice are permanently allocated to a particular user group

EXAMPLE: For emergency service usage, permanently reserved and always available in case of an emergency incident.

network topology: arrangement of network elements, aggregations of network elements, the relationship between the elements/aggregations, endpoints of connections (termination points), and transport entities (such as connections) that transport information between two or more termination points across the network topology

NOTE 1: See ETSI TR 121 905 [i.3].

NOTE 2: In an NFV implementation the network elements are replaced by network functions as explained in the ETSI ISG NFV set of specifications ETSI GS NFV 002 [8] and ETSI GS NFV 003 [9].

NOTE 3: For a formal set based definition of a topology, please see [i.8] and [i.9].

next generation vehicles: generic term for the following entities: drone, autonomous drone, autonomous vehicle, connected vehicle

next hop: neighbour/neighbour which has been selected to forward packets along the way to a particular destination

NOTE: As defined in IETF RFC 3753 [12].

Non-Access Stratum (NAS): that part of a cellular network that is not the Access Stratum

EXAMPLE: EPC core network and UE.

(N)-Port: binding of an EFCPM-instance to either an Application-Entity-Instance or RMT-instance in the layer above

NOTE: See ISO/IEC 7498-1 [6].

(N)-Port-id: identifier unambiguous within the scope of the two ends of a port

NOTE 1: The scope of this identifier is more restricted than a CEP-id.

NOTE 2: In practice a Port-id may have the same scope. However, they are still distinct namespaces.

NOTE 3: In the Internet, port-id = CEP-id = socket, see ISO/IEC 7498-1 [6].

(N)-Protocol: syntax of (N)-PDUs, and associated set of procedures, which specifies the behaviour between two (N)-PMs for the purpose of maintaining coordinated shared state (commonly known as communication peers)

NOTE 1: Commonly known as communication peers, see ISO/IEC 7498-1 [6].

NOTE 2: Protocol may be implemented in hardware and/or software.

(N)-Protocol-control-information PCI: portion of an (N)-PDU that is interpreted by the (N)-PM to maintain shared state of the protocol

NOTE: For a layer (N). See ISO/IEC 7498-1 [6].

(N)-Protocol-data-unit: unit of data exchange by (N)-PMs consisting of (N)-PCI and (N)-user-data

NOTE: For a layer (N). See ISO/IEC 7498-1 [6].

(N)-Protocol-machine: finite state machine that implements an (N)-protocol, which exchanges PDUs with a peer to maintain shared state with a corresponding (N)-PM, usually in another processing system

(N)-Relaying/multiplexing task: task within IPC-Process that performs multiplexing and/or relaying of PDUs

NOTE: There is one RMT in each IPC Process.

(N)-Service-data-unit: contiguous unit of data passed by an (N)-PM in an IPC API primitive whose integrity is to be maintained when delivered to a corresponding application protocol machine

NOTE: For a layer (N). See ISO/IEC 7498-1 [6].

N)-User-data: that portion of an (N)-PDU that is not interpreted and is not interpretable by the (N)-PM and is delivered transparently to its client, as an (N)-SDU. (N)-user-data may consist of part of, precisely one, or more than one (N)-SDU

NOTE 1: If more than one (N)-SDU, then SDUs in the (N)-user- data are delimited by the (N)-PCI.

NOTE 2: For a layer (N). See ISO/IEC 7498-1 [6].

piloting: driving, flying or aquatic steering of a vehicle

prefix: bit string that consists of some number of initial bits of an address

NOTE: As defined in IETF RFC 3753 [12].

processing system: hardware capable of supporting tasks that can coordinate with a "test and set" instruction (i.e. the tasks can all atomically reference the same memory)

NOTE: For a layer (N). See ISO/IEC 7498-1 [6].

RaaS: RAN as a Service.

rejoin: future access connection capability whereby a user can setup a logical connection to an access network, subsequently park the connection and then re-join the connection with the same or largely the same configuration at a later time without having to setup the connection again from scratch

NOTE: This type of connection is envisaged as being able to support one or more users.

route entry: entry for a specific destination (unicast or multicast) in the routing table

NOTE: As defined in IETF RFC 3753 [12].

Routing Locator Routing Locator (RLOC): actual location of an entity

NOTE 1: RLOC is the location of where the entity resides in the topology of the network.

NOTE 2: The binding EIDs and RLOCs enable an Entities to be located and reached, see IETF RFC 6830 [4], [14] and [15] 'where' in this context can have different values of graphical scope.

routing table: table where forwarding nodes keep information (including next hop) for various destinations

NOTE: As defined in IETF RFC 3753 [12].

service chaining: setup of a flow between two communicating peers for one or more users on a network that that operates across nominated SDN virtualised transmission entities and may include associated interconnected VNFs

subnet: logical group of connected network nodes

NOTE: In IP networks, nodes in a subnet share a common network mask (in IPV4) or a network prefix (in IPv6).

vehicle: entity used for transporting people, goods or sensors, including: those operating in ground-based environments (including rail-mounted and tunnelling), airborne or aquatic environments

NOTE: Where the term 'vehicle' is used within the present document, accompanying text should be supplied to specify the applicable set of environments.

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in ETSI TR 121 905 [i.3] and the following apply to scenarios that include mobile network architectures.

3GPP™	3rd Generation Participation Project
ACO	Automatic Cluster Optimization
ANO	Automatic Network Organization (Future SON)
AP	Access Point
AP	Application Process
AS	Access Stratum

BBU BaseBand Unit (of C-RAN)
CA Congestion Avoidance
CC Congestion Control
CE Customer Edge router

NOTE: A.k.a. CPE, customer premises equipment/edge Router.

CM Configuration Management (OAM)
CLC CooperativeLane Change of automated vehicles
CoCa Cooperative collision avoidance (CoCA) of connected automated vehicles
CP Control Plane
CPRI Common Public Radio Interface
C-RAN Cloud RAN
CTN Core Transport Network
CUPS Control and User Plane Separation
D2D Device to Device communication
DASH Dynamic Adaptive Streaming over HTTP (DASH)

NOTE: A.k.a. MPEG-DASH.

DC Dual Connectivity
DHCP Dynamic Host Configuration Protocol
DNS Domain Name Server
D-RAN Distributed RAN
DRB Data Radio Bearer
xDSL X-Digital Subscriber Line
DTP Data Transmission Protocol
DTCP Data Transmission Control Protocol
ECN Explicit Congestion Notification
EFCP Error and Flow Control Protocol
EFCPM Error and Flow Control Protocol Machine
EID Endpoint ID
EM Element Manager (OAM)
eMBB Enhanced MBB
eNB LTE evolved Node-B (Base Station)
EPC Evolved Packet Core
EPS Evolved Packet System (EUTRAN + EPC)
ETE End-To-End
EUTRAN Evolved UTRAN
FC Flow Control
FDPeC Flat Distributed Personal Cloud
FM Fault Management (OAM)
FMC Fixed Mobile Convergence
fps frames per second
GVE Generic Virtual Encapsulation
FTTx Fibre To The (x)
H-RAN Hybrid RAN
HSS Home Subscriber Server
HST High Speed Train
HTTP Hyper-Text Transport Protocol
ICIC Inter-Cell Interference Coordination

NOTE: Relevant to RAN SON.

IEEE Institute of Electrical and Electronic Engineering
IETF Internet Engineering Task Force
IMS IP Multimedia Sub-system
IoC Information object Class
IoT Internet-of-Things
IP Internet Protocol
IPC Inter-Processor Communications
ISD Inter-Site Distance

ISG Industry Specific Group

NOTE: Of ETSI SDO.

LISP Locator/ID Separation Protocol
MANO MANagement and Orchestration (of NFV framework)
MBB Mobile BroadBand
MBH Mobile Backhaul
MEC Mobile Edge Computing
MeNB Master eNB of a DC session
MME Mobility Management Entity
Msg Message
vMSO virtual Mobile Service Operator
MTP Motion To Photon

NOTE: Relevant to Video term.

NAS Non-Access Stratum
NE Network Element
NF Network Function
NFV Network Function Virtualisation
NGA Next Generation virtualisation Agent
NGC Next Generation virtualisation Controller
NGP Next Generation Protocols
NGMN Next Generation Mobile Network
NMS Network Management System
NSSF Network Slice Selection Function
NR New Radio
NS Name-Space

NOTE: Relevant to naming topic.

NS Network Service

NOTE: Relevant to Function Chain definition in ETSI GS MEC 003 [20].

OAM Operations, Administration and Management
OMC Operations and Maintenance Centre
OTT Over The Top (service)
PaaS Platform as a Service
P-Router Provider Router
PDU Protocol Data Unit
PE-Router Provider-Edge Router
PeCM Personal Content Management
PCI Physical Cell ID optimization algorithm (SON)
PCRF Policy, Charging & Rules Function
PDN Packet Data Network
P-GW Packet Gateway NE
P-GWc Packet Gateway, CP Functionality
P-GWu Packet Gateway, UP Functionality
Pk Packet
PLMN Public Land Mobile Network
PM Performance Management (OAM)
PM Protocol Machine (Protocols)
PSC Public Safety Communications
PUC Public Utility Communications
QoE Quality of Experience
QoS Quality of Service
RAN Radio Access Network
Rb Bit Rate
RF Radio Frequency
RINA Recursive Inter-Networking Architecture
RLOC Routing LOCator

RMT Relaying/Multiplexing Task
ROI Range Of Interest (video term)
RRU Remote Radio Unit

NOTE: Of C-RAN, a.k.a. Remote Radio Head - RRH.

SaaS Software as a Service
SBI South Bound Interface

NOTE: Relevant to OAM, from EM to NEs, NFV/SDN from VIM to VNFs.

URLLC Ultra-Reliable, Low Latency Communications

NOTE: See ETSI TR 138 913 [i.25].

SB-P South Bound interface Protocol
SCADA Scanning Control And Data Acquisition
SDN Software Defined Networking
SDU Service Data Unit
SeNB Slave eNB

NOTE: There are one or more Slave eNB's in a DC session.

S-GW Serving Gateway NE
S-GWc Serving Gateway CP Functionality
S-Gwu Serving Gateway UP Functionality
SON Self-Organizing Networks
TCP Transmission Control Protocol
UHD Ultra-High Definition (video standard)
UP User Plane
URL Universal Resource Locator
URLLC Ultra-Reliable, Low Latency Communications

NOTE: See ETSI TR 138 913 [i.25].

VaD Video data sharing for assisted and improved automated driving (VaD)
V2X Vehicle to another entity (X)
V2V Vehicle to Vehicle
V2I Vehicle to Infrastructure
VNF Virtual Network Function
VPN Virtual Private Network
VR Virtual Reality
XML eXtensible Mark-up Language

4 Overview

The Next Generation Protocols (NGP), ISG aims to review the future landscape of Internet Protocols, identify and document future requirements and trigger follow up activities to drive a vision of a considerably more efficient Internet that is far more attentive to user demand and more responsive whether towards humans, machines or things.

A measure of the success of NGP would be to remove historic sub-optimized IP protocol stacks and allow all next generation networks to inter-work in a way that accelerates a post-2020 connected world unencumbered by past developments.

The NGP ISG is foreseen as having a transitional nature that is a vehicle for the 5G community and other related communications markets to first gather their thoughts together and prepare the case for the Internet community's engagement in a complementary and synchronized modernization effort.

Therefore, NGP ISG aims to stimulate closer cooperation over standardization efforts for generational changes in communications and networking technology.

One of the biggest issues that should be addressed by NGP is the efficient operation of Internet Protocols over substantially heterogeneous networks in order to provide an acceptable End To End (ETE) user Quality of Experience QoE. In order to apply bounds to this problem the following domains are introduced:

1) Fixed and/or wireless and/or mobile access domain, known as the Access Stratum or AS, e.g. AS(Fixed), AS(Wireless), AS(Mobile).

2) Access Interconnection domain, which interconnects the AS with an operators Core Transport Network (CTN) commonly known as backhaul and sometimes also including front-haul for the case of Cloud-RAN (C-RAN).

3) An operator's core transport network domain, CTN, which includes the Non-Access Stratum (NAS) such as an Evolved Packet Core in the case of LTE.

4) CTN interconnection domain, which interconnects the CTN with other operators and/or PDNs.

The present document introduces the NGP, ISG view on key issues with today's Internet Protocol (IP) suite when operated so as to interconnect these domains.

In order to address the issues raised by the NGP, ISG, the present document introduces a reference set of scenarios that exemplify the current issues experienced in the operation of the existing IP suite for the NGP ISG to use in order to compare and contrast existing IP suite protocols with next generation IP suite protocol proposals.

The present document also lists example use cases that should be considered as typical for each scenario, but it does not introduce any new use cases but instead references existing use case definitions from standardization work in the next generation architecture and network standards market.

Each scenario is defined in terms of the following parts:

i) A model scenario architecture

ii) Model scenario description

iii) A description of the NGP agreed issues with the application of the IP suite to the scenario

iv) A list of applicable use cases

v) A list of targets to meet the next generation use cases

5 Issues to be addressed by the Scenarios

The NGP, ISG aims to address the following issues, as summarized in Table 1.

Table 1: NGP Key Issues

Issue ID	Issue Name	Issue Description
01	Addressing	Addressing needs to be scalable, per layer and separate from location and application name and instance
02	Security	Native IP is insecure. Many piecewise IP suite protocol "enhancements" have been added on to the suite but most are not part of a scalable, common security model and so do not always work well together and when used without care can cause notable issues with performance. e.g. the blanket application of HTTPS by some service providers is unnecessary and stifles mobile network optimization as all of the TCP payload is encrypted and tags that are used to optimize are hidden from the optimization algorithms. e.g. the TCP pseudo header is not a pure security integrity check and causes efficiency problems by merging layer functionality: a) Identity Adding Identity service and Management (logging/checking/validation, management) is essential in order to avoid many security vulnerabilities in the current IP suite. b) Location Adding low energy, scalable accuracy location information management is becoming essential for many services such as mobility, proximity and direction finding applications but is not present in most protocols today and only ever as an overlay which is usually inefficient. c) Authorization It is necessary in today's complex networks and internetworking to enable authorization capabilities to improve access and Network and Third Party Service provider authorization to be always 2-way and always secure. d) Accounting Needs to be included to support licence conditions and/or economic value to be apportioned. The solution needs to be scalable, traceable and cost effective. e) Auditing Essential for many licence conditions. Needs to always possible, easily controllable by multiple authorized entities/stakeholders and able to be checked by the user where appropriate. This issue group Includes Lawful intercept support. f) Authentication Required to include authenticating ID, Authorization and Messages, Message Content before exchanging secure/value based data.
03	Mobility	Mobile IP is limited in effective use for Mobile Cellular networks (slow to propagate and complex) and 3GPP's GTP although a pragmatic solution for small managed areas of scope, it is not efficient, particularly when operating multiple tunnels ETE and when operating multiple different QoS bearers at once. The mobility problem is exacerbated when operated with Dual Connectivity (multi-Cell) across multi-point Cells and with multiple different QoS bearers provided towards the same device.
04	Multi-Access Support. (including FMC)	IP does not currently accommodate a device that accesses multiple access networks at the same time as is becoming commonplace in for example: LTE and will be more common in the 5G era. (e.g. N x Access technologies from: Fixed, Wi-Fi, Cellular-RF, Cellular-mm Wave). This issue also includes Fixed Mobile Convergence in the core network across multiple access technologies.
05	Context Awareness	IP is not Context-Aware and so cannot respond explicitly to changing User Behaviour, Transport, Location and/or Situation (this includes mobility support issues).

Issue ID	Issue Name	Issue Description
06	Performance (including Content Enablement)	The current heterogeneous networks that are in operation today have several performance issues that are frequently experienced by users, such as: i) notable time to gain access to a particular network (fixed, Wi-Fi and/or cellular); ii) latency per requested transaction, session or call; iii) video issues: e.g. dropouts and restart. The scenarios for this issue provides examples off the current scenarios where these issues are experienced today. The scenarios is for this issue explain some of the problems that are related to how the IP-suite handles transport across heterogeneous pieces of an ETE communications path per transaction and/or session and/or call, e.g. TCPs built-in congestion control. Also, the lack of coordination between layers of context and congestion in order to drive any transmission issue avoidance and/or mitigation controls. Requirements from ultra-low latency use cases from different sectors need to be supported for next generation networks (i.e. automotive) whilst still supporting such simple communications transactions as page downloads in a timely manner. The performance issue includes the enablement of content in the network as follows: Requirements from video and content distribution, including Published and Private Video, Music and Documentation (books, documents, etc.). There is no established Mobile Content Management support in the IP suite that addresses such features as mobile edge caching control, for new entities such as Mobile Edge Computing Devices, this should be accommodated in the performance issue as part of the recommendations for improvement.
07	Network Virtualisation	SDN is becoming established. However, Network Functional Virtualisation (NFV), whilst defined in ETSI-ISG-NFV by procedure and interface(s) does not have an agreed management protocol for orchestration topology and/or to manage vNF scale & scope within the MTC. Management of future networks with integrated virtualisation control is now essential to support such emerging features as Network Slicing and apply network SON algorithms such as Automatic Cluster Optimization (ACO). (See note).
08	IoT support	Scenarios that highlight requirements from the Internet-Of-Things for NGP.
09	Energy Efficiency	Scenarios that highlight the requirements for increased energy efficiency within the global ICT sector for NGP.
10	e-Commerce	Scenarios that highlight requirements from eCommerce for NGP.
11	MEC	Mobile Edge Computing is seen as a critical feature of Next Generation communications systems for low latency and service acceleration that cannot be achieved with current systems. This scenarios highlights the requirements from MEC for NGP.
12	Mission Critical Services: Public Safety and Public Utility Communications	Historically there have been separate dedicated communications infrastructure and protocols for the support of Mission Critical services for both Public Utility Communications (PUC) and Public Safety Communications (PSC). However, with the mounting costs of communications infrastructure and maintenance, many countries are considering common communications infrastructure to support public communications and PUC and PSC. Common infrastructure support is envisaged through the adoption of techniques such as network slicing and other much specified but little implemented techniques such as RACH classes, access priority, MLPP and eMLPP. This scenario area is introduced where the supporting NGP architecture will also need to support infrastructure bearing PUC and PSC traffic. Also, the NGP protocol architecture will need to inherently support the additional services that PUC and PSC users demand such as Group Calling and extremely efficient robust voice and data services such as PTT and Group Messaging under emergency and/or event based ad-hoc communications conditions. No new scenarios are identified in the present document, only the requirement references for NGP protocols to support are introduced. Requirements applicable to NGP for Mission Critical Services support will be referenced to existing 3GPP documents in the forthcoming NGP requirements document.

Issue ID	Issue Name	Issue Description
13	Autonomous and Connected Drones and Vehicles	With the recent interest in the employment of autonomous drones for surveying monitoring, delivery and reconnaissance and the growth of interest in Autonomous Vehicles and Connected Vehicles over the last few years, the supporting communications network protocols for the next generation will need to accommodate support for these use cases. This section addresses scenarios where such use cases will become prevalent in the next few years.
14	Ultra-Reliable Low Latency Communications	This section introduces scenarios that have been defined for next generation protocol usage in the 3GPP 5G scope that require Ultra-Reliable communications and/or Low Latency communications.
NOTE:	(From NGP#02) since it is envisaged that next generation networks will be notably virtualised, this scenario needs to highlight: i) Management and Organization scenarios from the traditional OMC approach; and ii) the emerging Virtualised, network operational approach).	

6 Model References

6.0 Introduction

Clause 6 provides reference network architecture and protocol models for the purpose of highlighting the use of the currently IETF defined IP protocols in supporting current networks.

6.1 LTE Mobile Network Model

This clause provides a reference model for a 3GPP LTE network and highlights its use of the IP protocol suite.

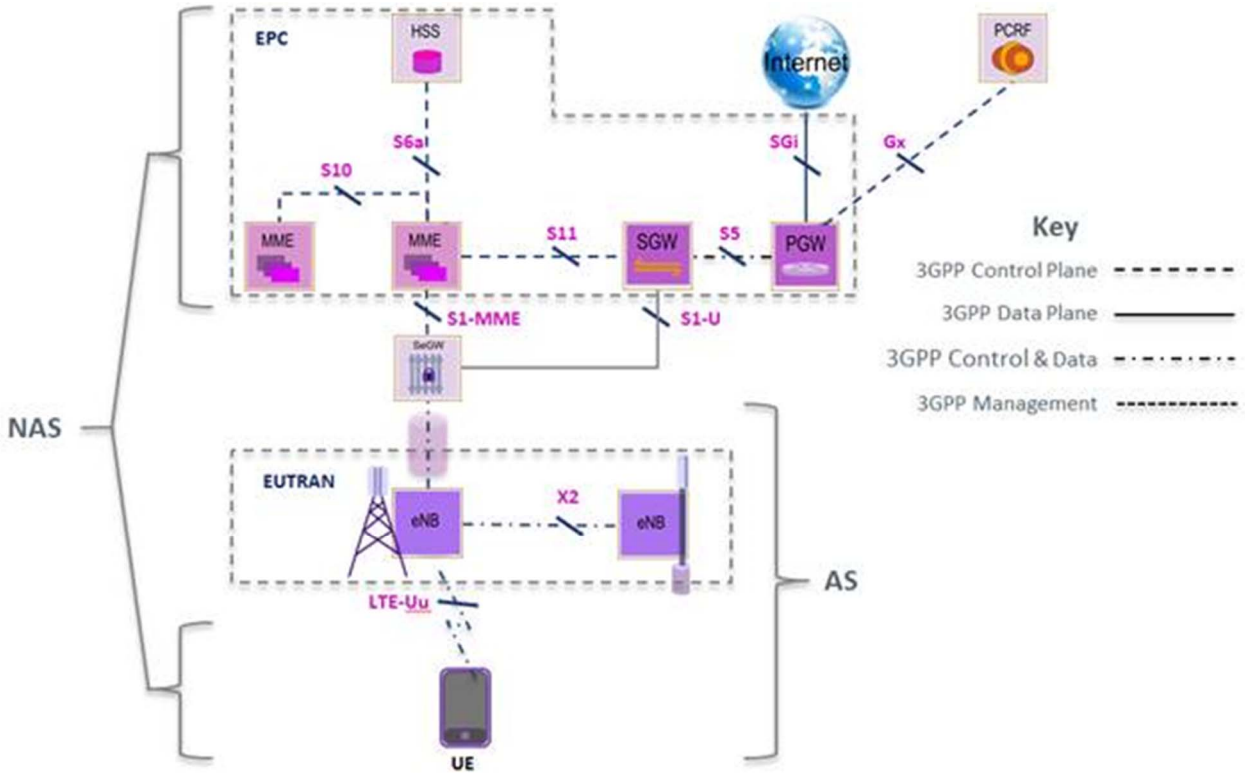


Figure 1: 3GPP LTE Rel-12 Architecture

Figure 1 illustrates the currently widely deployed 3GPP LTE Release 12 architecture. The protocols that support this architecture are illustrated for the 3GPP User Plane (UP) in Figure 2.

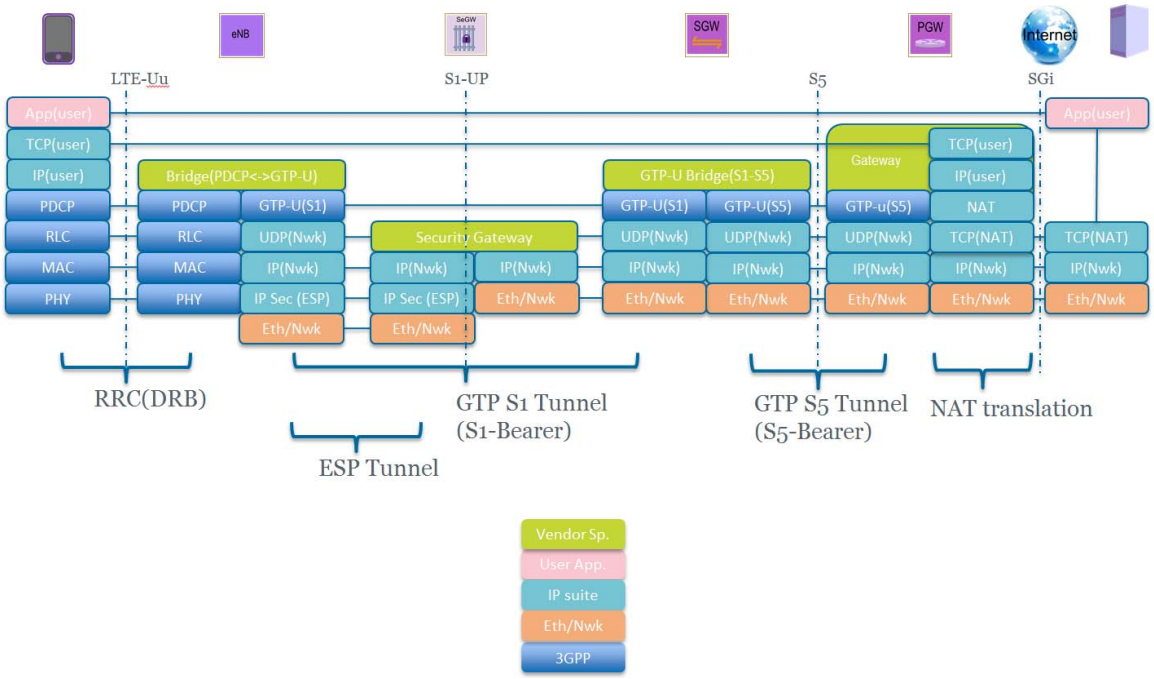


Figure 2: 3GPP LTE Rel-12 UP Protocols

Figure 2 colour codes the IP suite protocols into the following categories for illustrative purposes:

i) 3GPP defined protocols, illustrated in "blue".

ii) Proprietary bridging and gateway functionality at each intermediate node that is used to integrate the ETE UP path, illustrated in "green".

iii) Network interconnecting layer 2 protocol, illustrated as Ethernet which is true in most cases but may also be other L2 transmission protocols, illustrated in "orange".

iv) The ETE application protocols that are using this ETE UP protocol stack/path to communicate, illustrated in "pink".

Figure 5 illustrates the standardized protocols required to realize a 3GPP UP over a standardized 3GPP, Release 12 LTE architecture. Figure 2 additionally includes some of the key practical protocol considerations that need to be addressed in any realistic commercial mobile network LTE implementation, as follows:

i) The adoption of a Security Gateway between the eNB and the S-GW which secures the user plane when the network includes eNB's that are physically remote from the rest of the network equipment. Although this is usually universally adopted as good practice in many LTE networks today.

ii) NAT functionality deployed at the P-GW on a per PDN/APN breakout basis to administer intranet <-> Extranet address mapping between the operator network address range allocated to the mobile users within the private IP operator network and the external, public, static IP address range of the operator (usually very small range for the internet) and public TCP ports used to proxy the mobile user in order to communicate with the internet or other external PDN(s).

Figure 5 also illustrates the key logical bindings that form the ETE, LTE, UP as follows:

i) Data Radio Bearers or **DRB**, one for each different Traffic Flow Template (TFT) currently in use by the user over their Radio Resource Connection (RRC).

ii) The Ipsec, **ESP Tunnel** between the eNB and the Security Gateway for the user plane.

iii) The GTP S1 Tunnels which carry **S1-Bearers** between the eNB and the S-GW.

iv) The GTP S5 Tunnels which carry **S5-Bearers** between the S-GW and the P-GW.

The protocols that support the LTE architecture of Figure 1 are illustrated for the 3GPP Control Plane (CP) in Figure 3.

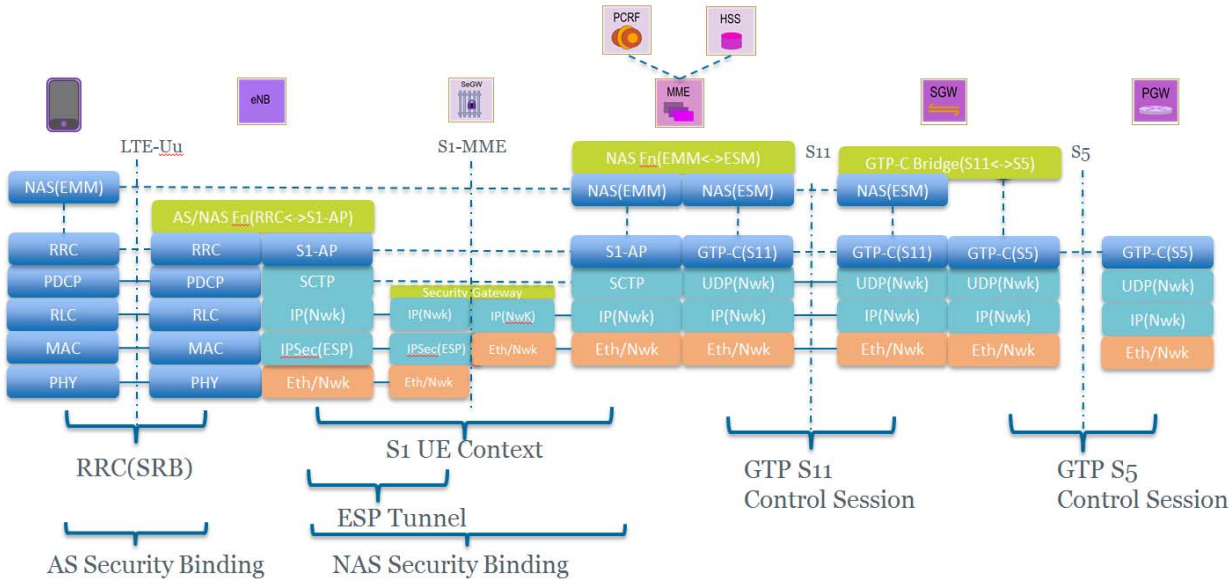


Figure 3: 3GPP LTE Rel-12 CP Protocols

Figure 3 is colour coded as per the key in Figure 2. Figure 3 illustrates the key logical bindings that form the ETE, LTE, CP as follows:

i) Signalling Radio Bearer (**SRB**) over the LTE-Uu air interface between the User Equipment (UE) device and the evolved Node-B (eNB), base station, over the users Radio Resource Connection (RRC).

ii) S1 signalling connection between the eNB and the MME as an **S1 UE Context.**

iii) S11 signalling connection between the MME and the S-GW as a GTP-C, CP, **S11 Session.**

iv) S5 signalling connection between the S-GW and the P-GW as a GTP-C, CP, **S5 Session.**

It is important to note that in establishing the ETE control Plane for a user, security bindings also need to be setup, as follows:

i) Over the Access Stratum (AS), there is an **AS security Binding** established between the UE and the eNB.

ii) Over the Non-Access Stratum (NAS), there is a **NAS Security Binding** established between the UE and the MME.

Figure 3 illustrates the standardized protocols required to realize a 3GPP CP over a standardized 3GPP, Release 12 LTE architecture. Figure 3 additionally includes some of the key practical protocol considerations that need to be addressed in any realistic commercial mobile network LTE implementation, as follows:

i) The adoption of a Security Gateway between the eNB and the S-GW which secures the IP transport for the control plane when the network includes eNB's that are physically remote from the rest of the network equipment. Although this is usually universally adopted as good practice in many LTE networks today.

It is also of noted that other signalling interfaces such as the S6a interface towards the Home Subscriber Server (HSS) (for SIM subscriber record and location updating and access authentication), the Gx (for PCRF policy charging and rules function management) and the Ga interface for charging between the LTE Nes and the Charging Gateway all operate DIAMTER/GTP-C/UDP/IP for support of their signalling procedures.

6.2 L2 and L3 VPN services

6.2.0 Introduction

Clause 6.2 introduces the most popular models used by service providers to deliver scalable Layer 2 and Layer 3 VPN services to their customers.

6.2.1 MPLS/BGP Layer 3 Virtual Private Networks

This clause provides a reference model for Layer 3 VPN network services (see IETF RFC 4364 [10]) and highlights its use of the IP protocol suite.

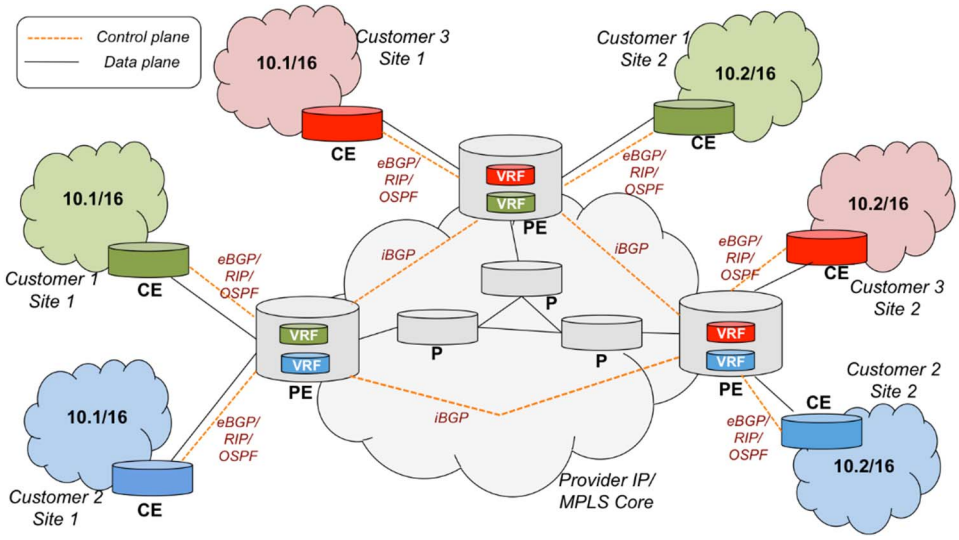


Figure 4: L3VPN model from IETF RFC 4364 [10]

Figure 4 illustrates an example of the L3VPN model from IETF RFC 4364 [10], showing a single service provider offering L3 VPN services to three different customers, each one with two sites. Customers can use overlapping address spaces, which are isolated in the service provider IP/MPLS core using a combination of Multiprotocol BGP (MBGP) in the control plane and MPLS in the data plane.

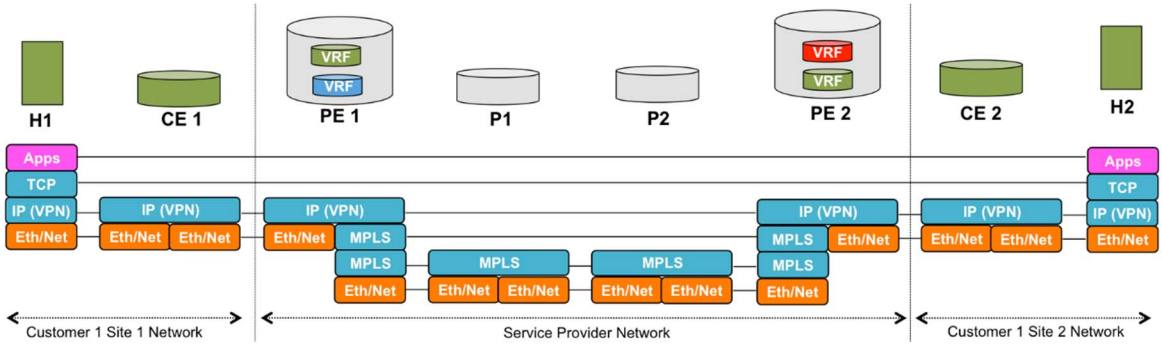


Figure 5: IETF RFC 4364 [10] Layer 3 VPN data plane

Figure 5 colour codes the IP suite protocols into the categories described in clause 6.1. The figure shows the user plane of a BGP/MPLS VPN service, with a service provider network connecting two sites of a customer. The layer 3 VPN is the IP layer that is "floating" on top of the provider's MPLS core. Each PE knows how to map customer traffic to a VRF (associating a data link layer "circuit" to a VRF). To do that it uses two MPLS labels: one that identifies the VPN and another ones that identifies the LSP between Pes through which the VPN traffic is forwarded.

The protocols that support the L3 VPN model of Figure 1 are illustrated for the Control Plane (CP) in Figure 6. Figure 3 is colour coded as per the key in Figure 2. Figure 6 illustrates the main protocols used in the control plane of the customer and provider networks. Starting at the customer network, the CE runs an eBGP session with the PE in order to exchange routes. Focusing on the provider network, two main tasks for the control plane can be distinguished:

- the exchange of VPN routes; and
- the exchange of label information related to the state of the MPLS LSPs.

The latter task is accomplished by the regular MPLS operation procedures. Pes use MBGP to exchange VPN route information.

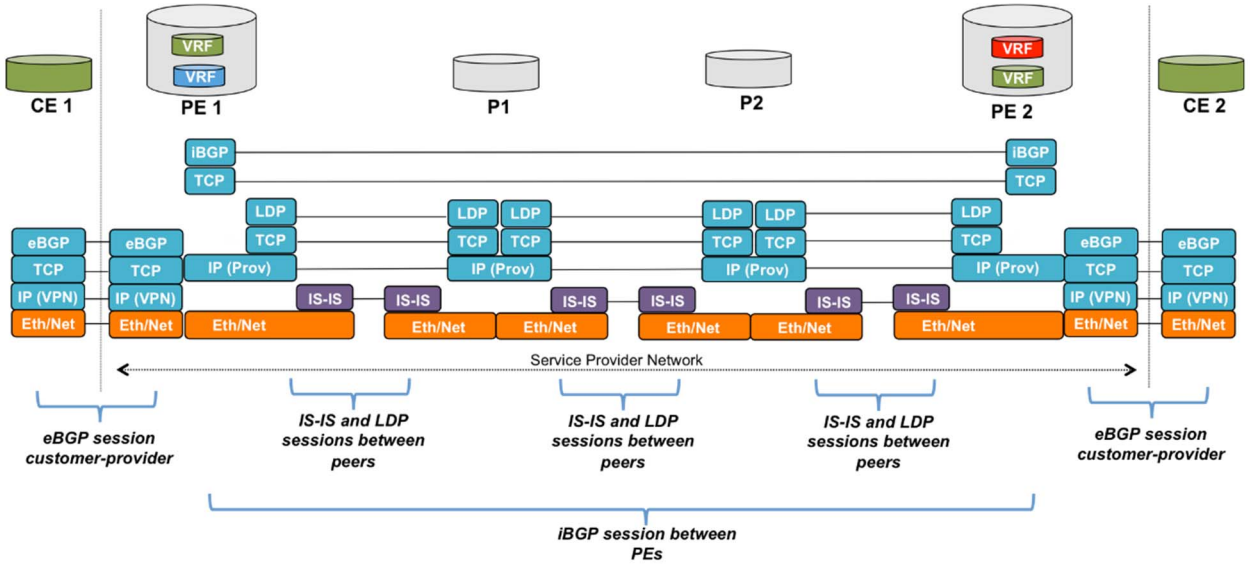


Figure 6: IETF RFC 4364 [10] Layer 3 VPN control plane

6.2.2 VPLS, Virtual Private Line Services and Ethernet-VPN

This clause provides a reference model for Layer 2 VPN network services (multi-point to multi-point) over an IP/MPLS network, and highlights its use of the IP protocol suite. Two options are possible for the VPLS control plane: BGP (IETF RFC 4761 [11]) and LFP (IETF RFC 4762 [i.15]). Figure 7 illustrates an example of the VPLS model from IETF RFC 4761 [11] and IETF RFC 4762 [i.15]. The model is similar to the L3VPN service, with the differences that it is a Layer 2 service and that address learning is performed by VPLS instances in each PE at the data plane (following the same behaviour as a layer 2 bridge).

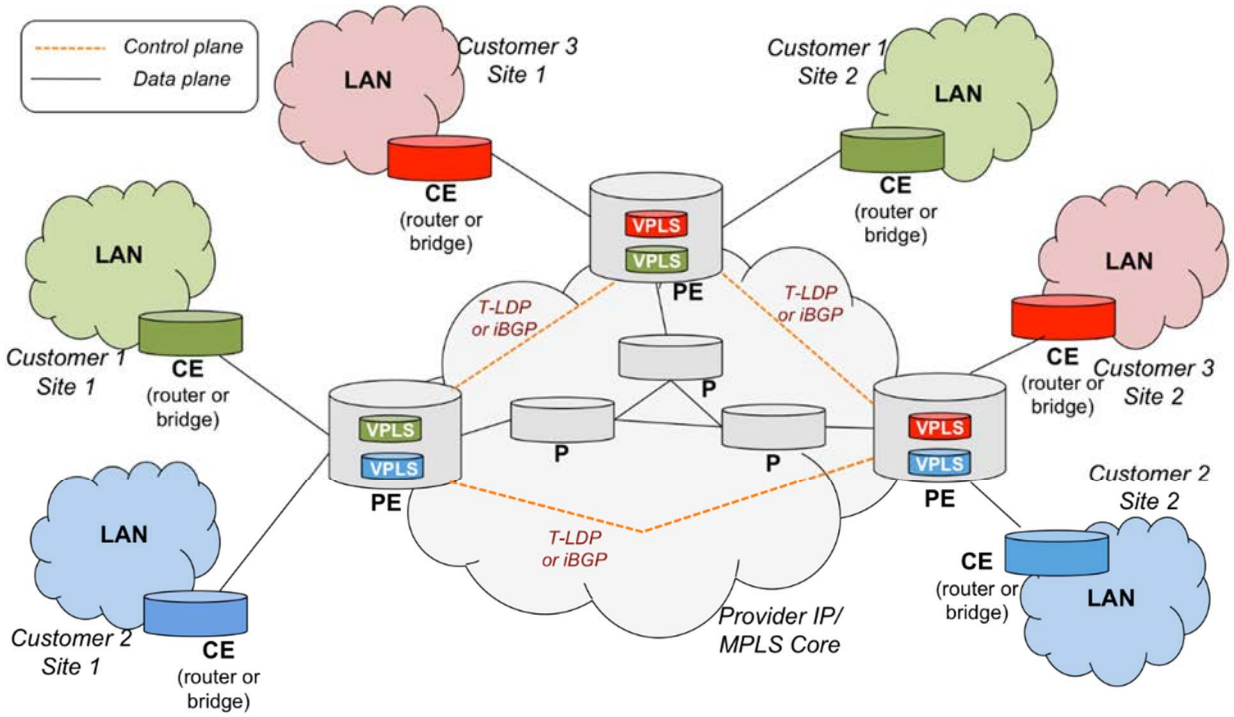


Figure 7: IETF RFC 4761 [11] and IETF RFC 4762 [i.15] VPLS architecture (CE directly attached to PE)

Figure 8 colour codes the IP suite protocols into the categories described in clause 6.1. The figure shows the user plane of a VPLS service, with a service provider network connecting two sites of a customer. Each PE knows how to map customer traffic to a VPLS instance (associating an Ethernet port or VLAN to a VPLS instance). Figure 9 is colour coded as per the key in Figure 8. Figure 8 illustrates the main protocols used in the control plane of the provider network. LSPs between Pes are established using the same procedures as in the L3VPN case. The Pseudo Wire mesh can be created through the use of MBGP or a mesh of T-LDP sessions between Pes.

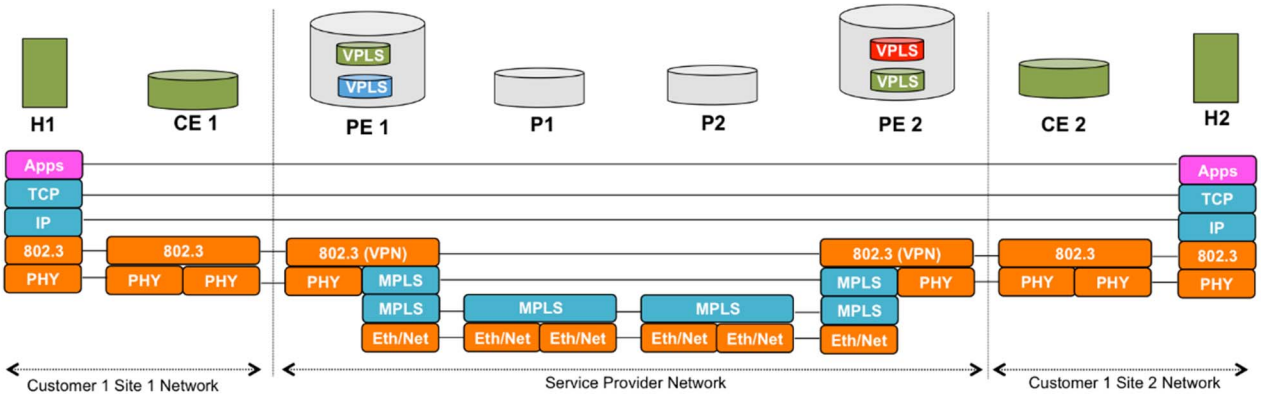


Figure 8: IETF RFC 4761 [11] and IETF RFC 4762 [i.15] VPLS data plane (CE directly attached to PE)

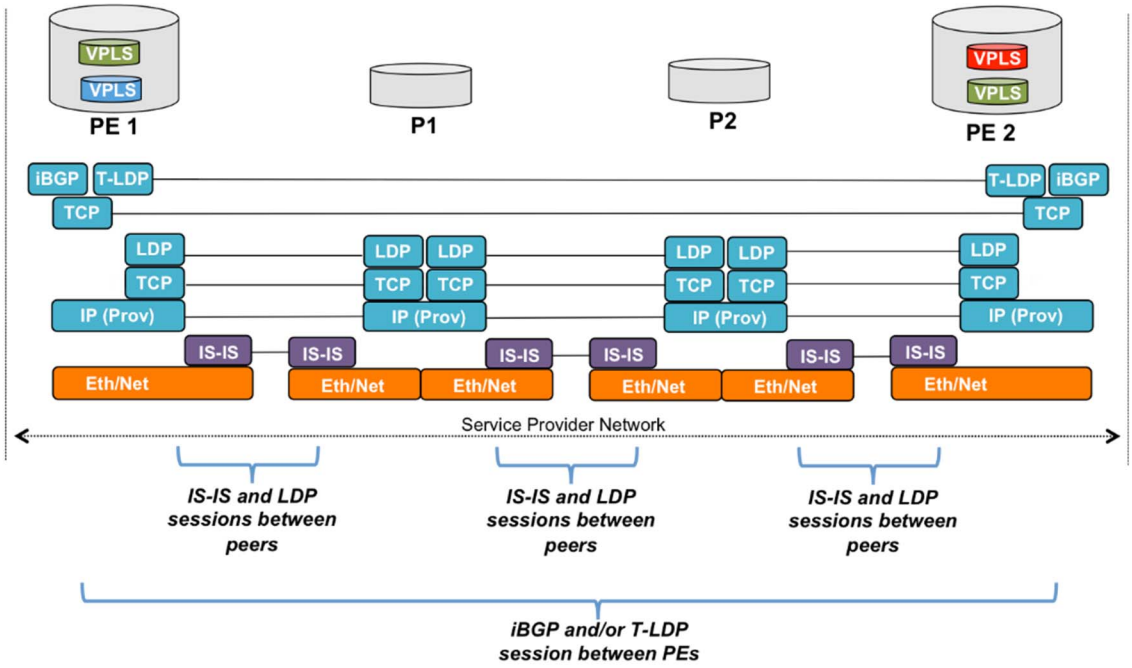


Figure 9: IETF RFC 4761 [11] and IETF RFC 4762 [i.15] VPLS control plane (CE directly attached to PE)

With the flat architecture depicted in Figure 7 it is hard for service providers to scale networks with large numbers of Pes (due to the requirements for a full mesh of Pseudo Wires). Hierarchical VPLS (H-VPLS) mitigates this issue by creating a hierarchy with different types of PE following a hub/spoke model, in which spoke Ces are directly attached to a hub, and the pseudo-wire mesh is only between hubs. H-VPLS still has scalability issues since the customer and the provider forwarding scopes are not isolated (hub Pes see all customer addresses). PBB-VPLS has been proposed to mitigate this issue (IETF RFC 7041 [i.23]). In PBB-VPLS traffic of customer VPLS instances is multiplexed into one or more Backbone VPLS instances using the capabilities provided by PBB (IEEE 802.1Q [26]), as shown in Figure 7.

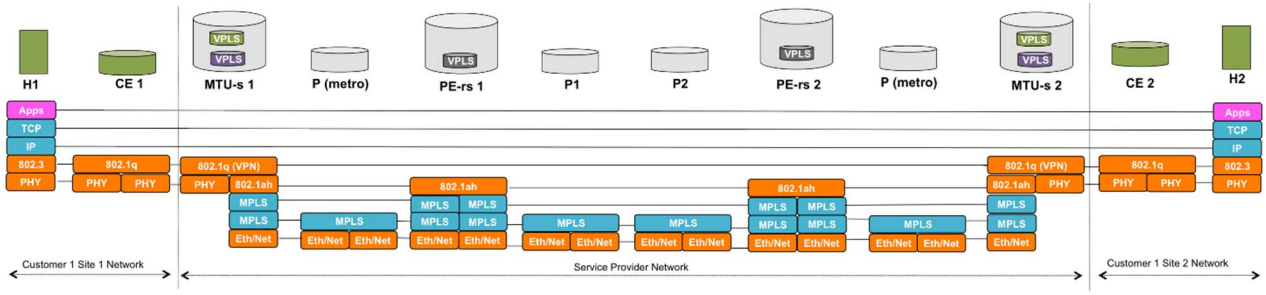


Figure 10: Example of PBB VPLS data plane

Ethernet VPN (EVPN) is the IETF response to limitations in the operation of VPLS, and follows similar operational procedures as L3VPN services, performing MAC address learning in the control plane. EVPN defines a single control plane (MBGP) with options for three different data planes:

- EVPN over MPLS;
- EVPN over PBB over MPLS (for isolating forwarding scopes and achieve higher scalability); and
- EVPN over NVO3 (to allow for L2VPNs over L3 without the need of MPLS).

6.3 All IP Core Network Model

This clause provides a reference model for all IP core networking, as Figure 11 shows. The essential point of the model is to allow the core network to carry services from various kinds of access networks such as enterprise VPNs, mobile access, PSTN, Broad Band, and Internet-of-Things, etc.

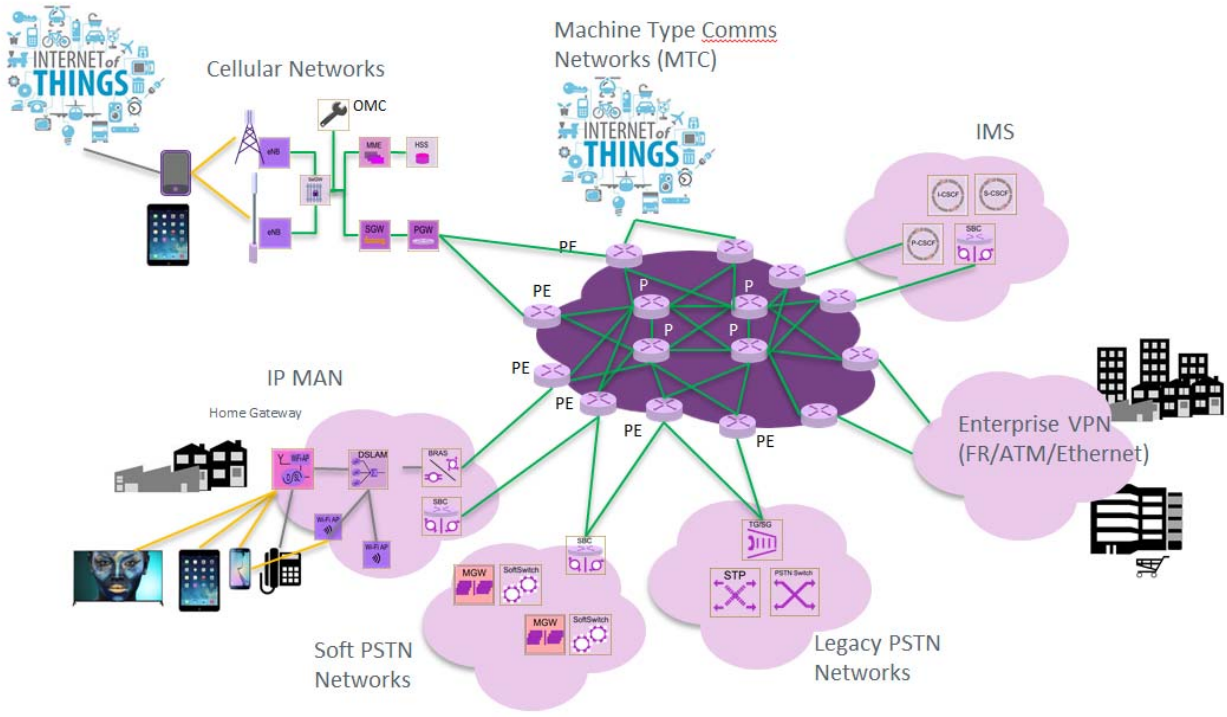


Figure 11: Carrying Multiple Services Using an All-IP Core Network

To connect with various access networks, the IP core network needs to provide different interfaces for communication. As illustrated in Figure 11, there are several common kinds of interfaces as detailed for the following interconnecting network types:

i) **Mobile Cellular Network:** mobile networks usually use different interfaces for different services such as voice communication and data communication. One specific interface usually implies some specific L2 or L3 technology, such as ATM, Frame-Relay and GTP-IP-Tunnelling, etc.

Currently for LTE (illustrated), the cellular network connects directly to the IP-Core or via a Security Gateway (SeG).

For legacy GPRS/EDGE networks the cellular network connects via Frame Relay and/or IP.

For legacy UMTS packet GPRS service the cellular network connects via ATM and/or IP.

For legacy UMTS circuit service, the cellular network connects via ATM and/or IP:

ii) Both Legacy and Soft-Switch based PSTN interfaces to the IP core need to be supported for interconnection of conventional legacy networks. Thus, the IP core should be able to emulate POTS services through IP transmission. For example, Pseudo-Wire is often employed for such a purpose.

iii) Native IP interfaces to the access network connect with the IP core directly through IP transmission. Metro and Enterprise VPN networks usually connect with the IP core through native interfaces.

iv) Today IP- MAN networks also usually support DSL home broadband interconnection and/or Cable domestic subscriber connection.

v) IMS networks are also connected to the IP core for the support of advanced packet voice services either via the Cellular network or as direct packet voice call support for enterprise business systems or for the relay of traditional voice services via packet means.

vi) IoT networks are either IP based natively in which case they connect natively to the IP core or they are interworked or they are relayed via Cellular networks either directly with a cellular interface at the Sensor(s) and/or Actuator(s) or via a mobile device acting as a gateway.

Many IoT interconnect options use GWs to support interworking to the IP core from BLE, Wi-FI™ and other IoT specific interfaces such as SIGFOX™ and LoRA™.

To fulfil the IP connected vision depicted in Figure 11, the IP core network approach is illustrated in Figure 12.

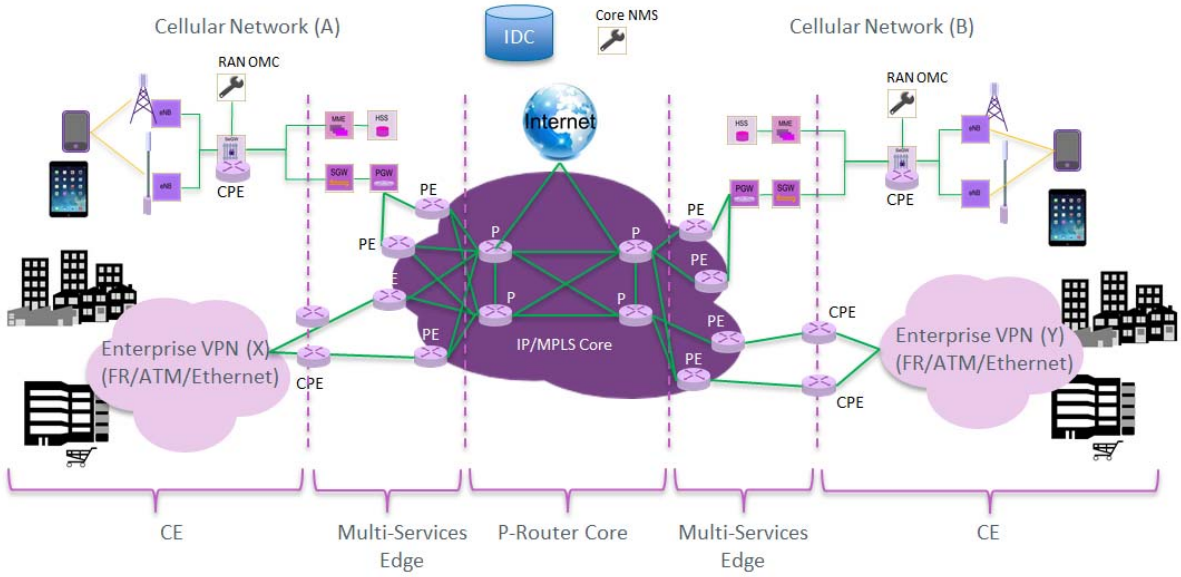


Figure 12: IP Core Network Structure

The IP core network needs to support both IP forwarding and MPLS forwarding. In modern devices, MPLS and IP are usually available on the same platform. IP forwarding is used mostly for broadband traffic. Services like PSTN, mobile access, and enterprise VPN which usually need more support on QoS, high availability, and security, etc., typically use MPLS VPN to guarantee quality and isolate different. For an introduction to IP VPN services see clause 6.2.

The IP core network is provided by a mesh of inter-connected Provider or (P) Routers that operate MPLS and are physically connected using typically MPLS over optical transmission.

At the edge of the IP/MPLS, IP-core network, there are Provider Edge (PE) Router nodes where multiple services are converged. These routers typically support many different types of transmission input, including optical, wired Ethernet, SDH, PDH, ATM and FR interfaces.

At the edge of each provider edge network site, Customer Edge (CE) Router nodes are deployed to connect customer equipment to the multi-service PE routers. These CE-Routers deliver service data to the relevant PE nodes in the form of IP/MPLS transit streams.

6.4 NFV Reference Model

The ETSI standards have introduced a 'reference architecture framework' model for Network Function Virtualisation (NFV) which is illustrated in Figure 13.

The NFV model is expected to be interconnected using IP as the base networking protocol and currently the control interfaces are all supported by TCP or SNMP operated over IP.

This framework is anticipated to be adopted widely in the next generation of networks. As such, the model is included in the present document for reference purposes and the NGP requirements are included in clause 8.7.

The key specifications for NFV are: Use Cases: ETSI GS NFV 001 [16], Architectural Framework: ETSI GS NFV 002 [8], Management and Orchestration: ETSI GS NFV 006 [i.40].

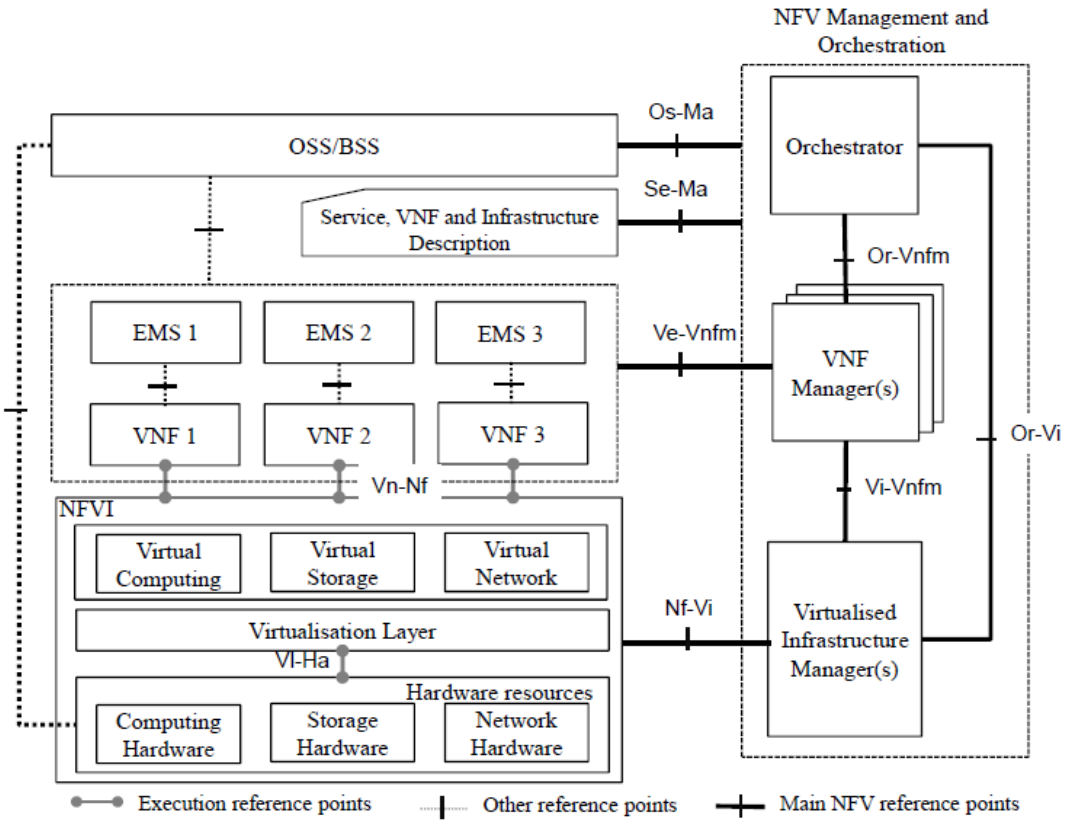


Figure 13: NFV Reference Architecture Framework (extracted from ETSI GS NFV 002 [8])

6.5 MEC Reference Model

ETSI standards have introduced a 'reference architecture framework' model for Mobile Edge Computing (MEC) which is illustrated in Figure 14.

The MEC model is expected to be interconnected using IP as the base networking protocol towards the core network and cellular protocols towards the Radio Access Networks (RAN) it supports, with IP network control interfaces operating TCP for transmission control.

The framework is introduced to support service and content acceleration and optimization at the edge of heterogeneous networks, particularly Radio connected networks where the radio access network is often the lowest performance of the ETE user communications path.

This framework is anticipated to be adopted widely in the next generation of networks. As such, the model is included in the present document for reference purposes and the NGP requirements for MEC are covered in clause 8.6.

The key specifications for MEC are:

- Service Requirements, ETSI GS MEC-IEG 004 [21];
- Architectural Framework, ETSI GS MEC 003 [20]; and
- Technical Requirements, ETSI GS MEC 001 [19].

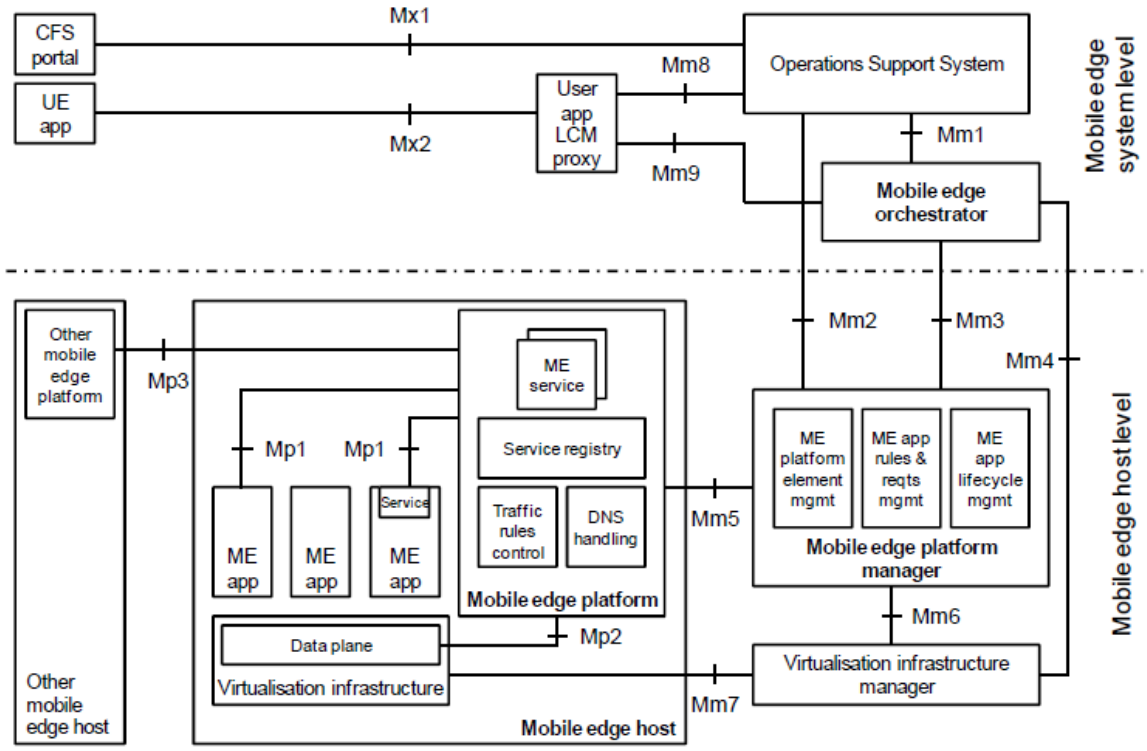


Figure 14: MEC Reference Architecture Framework (extracted from ETSI GS MEC 003 [20])

7 Referenced Use Cases

This clause lists the use cases that are used in the present document.

The Use cases are referenced from 3GPP TR 22.891 [i.1] which addresses input from other standards and research bodies such as [1] the NGMN Whitepaper on 5G and the 5GPP whitepaper output covered in [i.7], [i.6], [i.4] and [i.5].

The detailed parameterization of the use cases is defined in Annex A.

Table 2: Use Case References

UC Ref	Use Case	Use Case Group	NGP Scenarios Applicable
1	Ultra-Reliable Communications (includes URLLC IoT)	Capability	06
2	Network Slicing	Capability	05
4	Migration of services from earlier generations	Legacy	02
5	Mobile broadband for indoor scenario	MobileBB	02, 03, 06
6	Mobile broadband for hotspots scenario	MobileBB	02, 03, 05
7	On-demand Networking	MobileBB	02, 03, 05
8	Flexible Application Traffic Routing	Capability	05
9	Flexibility and Scalability	Capability	05
10	Mobile broadband services with seamless wide-area coverage	MobileBB	02, 05
11	Virtual Presence, Includes 360 degree Video control and 4K, 8K streaming 2D and 3D	MobileBB	06
12	Connectivity for drones	New	05, 06
16	Coexistence with legacy systems	Legacy	02, 05
18	Remote Control (includes High Speed Video relay)	Service	06
23	Access from less trusted networks	Legacy	02
26	Best Connection per traffic Type	Capability	05
27	Multi Access network integration	New	02, 03, 05
28	Multiple RAT connectivity and RAT selection	New	02, 03, 05
29	Higher User Mobility	Environ	03
30	Connectivity Everywhere	Environ	03, 05
32	Improvement of network capabilities for vehicular case	Environ	02, 03
33	Connected vehicles	Environ	02, 03
34	Mobility on demand	Environ	03
35	Context Awareness to support network elasticity	Environ	03
36	In-Network Caching	Capability	05, 06
37	Routing Path Optimization	Capability	03, 05, 06
38	ICN Content Retrieval	Capability	05, 06
42	Low mobility devices	IoT	03
47	SMARTER Service Continuity	Environment	03, 05, 06
48	Provision of essential services for very low-ARPU areas	Legacy	02
51	Network enhancements to support scalability and automation	Capability	02, 03, 05, 06

8 Scenarios

8.1 Addressing

8.1.0 Introduction

This scenario is described so as to identify issues for NGP with the current IP based addressing used for internet networking and transmission. All of the issues should be considered for current and evolving user access networks with respect to NGP. All proposed NGP solutions should demonstrate the capability to solve the identified issues introduced in this scenario.

8.1.1 Model Architecture

This scenario description firstly introduces a basic model for addressing scenario for NGP. This model is illustrated in Figure 15. This is a model that a User Equipment (UE) communicates with a peer device. The UE is moving across different EPC.

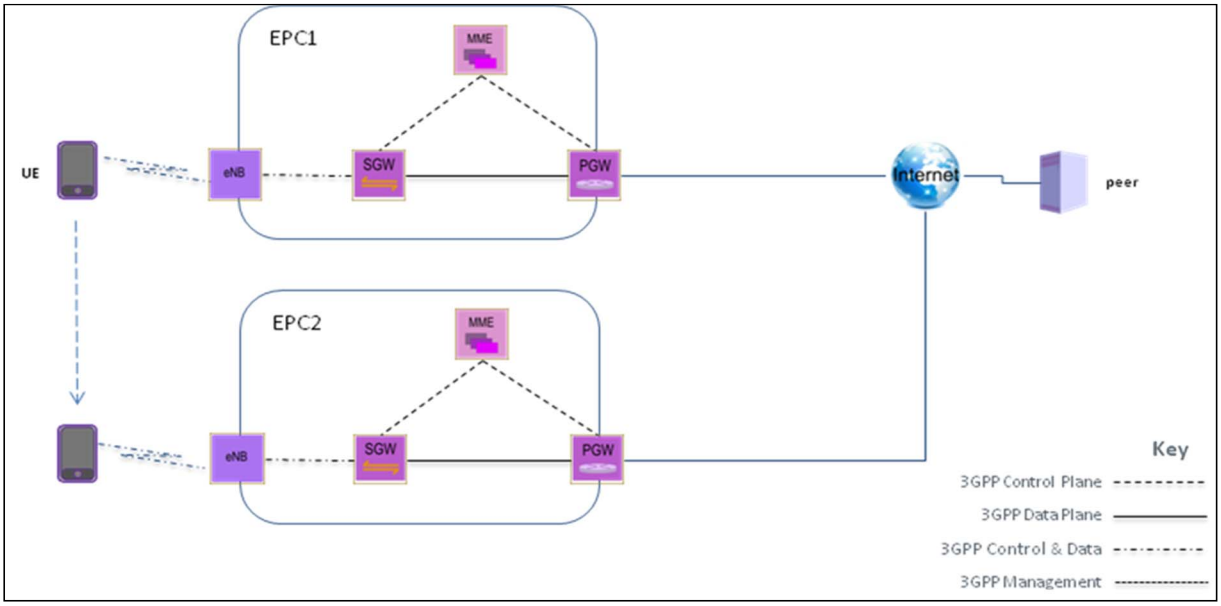


Figure 15: Model for UE moving across different EPC

The 2nd model is illustrated in Figure 16. This is a model that a UE communicates with a peer device. The UE is moving across different access network, i.e. from an LTE network to a Wi-Fi™ network.

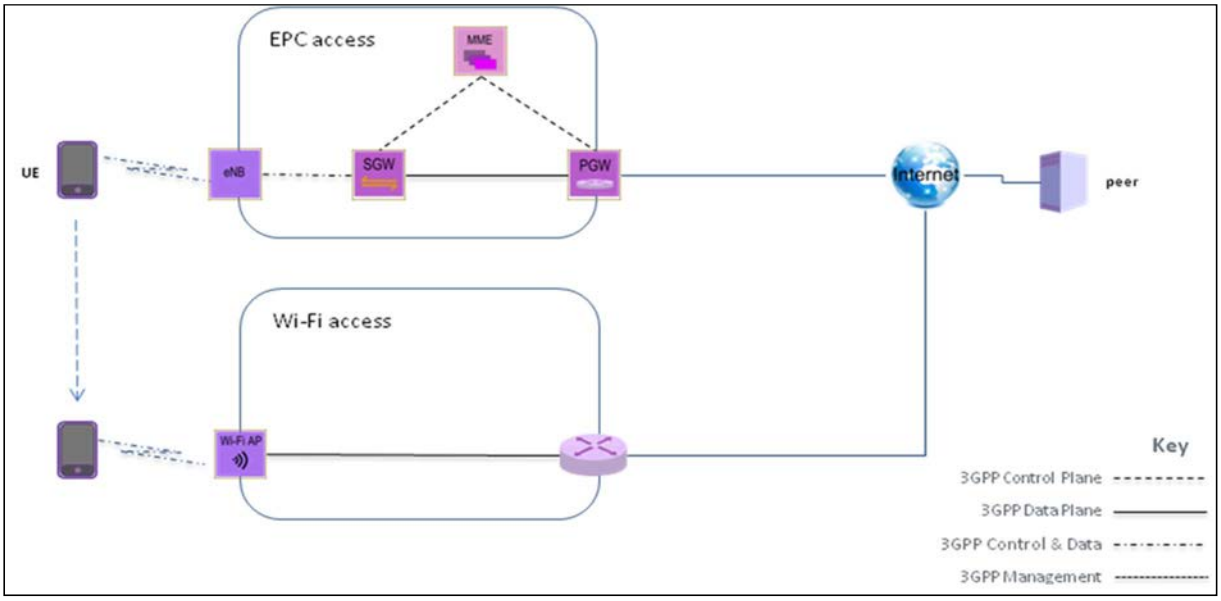


Figure 16: Model for UE moving across different access network

Figure 17 shows the 3rd model. This is a model that a customer wants to provide multi-homing with or without load balancing. The multi-homing server in customer network and is connected with two service providers. The customer network address space is assigned by one of its connected provider (Provider1).

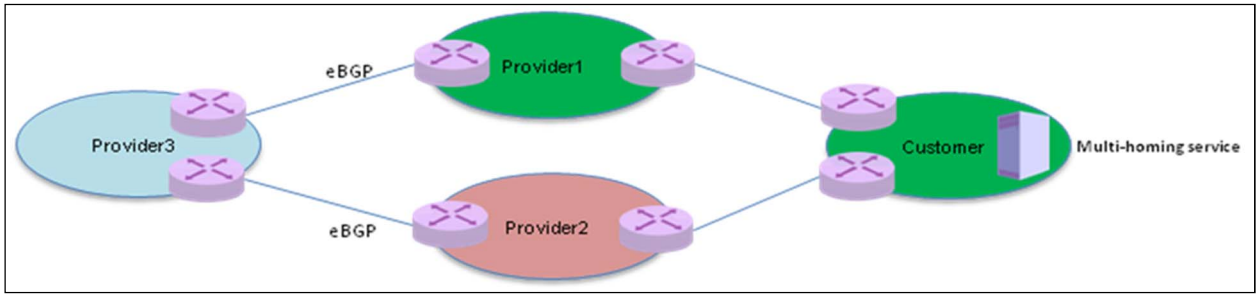


Figure 17: Model for multi-homing with provider assigned address

The next mode is shown in Figure 18. This is a model similar to Figure 3, but customer network address space is provider independent and not assigned by any of its provider.

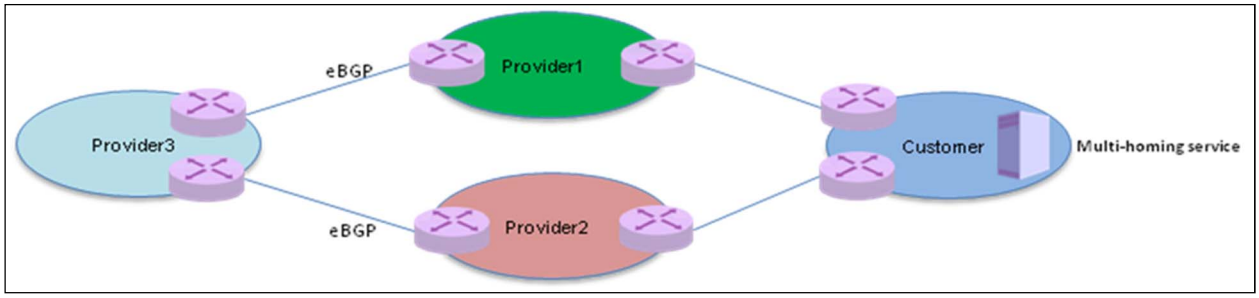


Figure 18: Model for multi-homing with provider independent address

8.1.2 Scenario Description

8.1.2.0 Introduction

This clause further describes the details in these scenarios.

Above models are for two typical scenarios to demonstrate the issues in the current IP based addressing system in internet:

- Mobile communication
- Multi-homing and load balancing

8.1.2.1 Scenarios for mobile communication

- IP address type for end-user

All end-user devices, including mobile user equipment (UE), should have an IP address assigned statically or by DHCP before connecting to the internet. The latter being the preferred solution. IP traffic is becoming more and more dominant in the in the wireless markets. For the foreseeable future, it is likely that most wireless traffic is IP based including traditional voice, except where operated in legacy networks.

In terms of the source or owner of IP address, there are two types of IP address an end-user device can obtain. One is provider assigned, and another is provider independent. Provider assigned IP address is assigned by a service provider when the end-user device is allowed to attach to the provider's network to get service. This address is normally from the address blocks the service provider owns. The provider independent address is directly allocated to an end-user organization by a Regional Internet Registry (RIR). Currently only Ipv6 address is available. In April 2009 RIPE (Réseaux IP Européens) accepted a policy proposal to assign Ipv6 provider-independent Ipv6 prefixes. Assignments are taken from the address range `2001:678::/29` and have a minimum size of a `/48` prefix.

- Provider assigned IP address

Using provider assigned IP address is the most popular way in the world for mobile service currently. It has many benefits for service provider in terms of technology maturity, business model and operational cost. i.e. service provider could assign private IP address and do NAT at Gateway, this is critical to some countries and service providers who should serve huge population but do not own enough address blocks, especially for Ipv4.

For wireless mobile communication, the models introduced in clause 8.1 are based on 4G architecture described in 3GPP. After the attachment, the UE should obtain IP address from a local IP address administrative entity. The local IP address administrative entity is a device which can provide the IP address assignment, such as DHCP functionality; it is normally at P-GW in 4G and LTE. This IP address can be either public or private, and is only valid in the limited administrative domain. If the mobile device is moving to different administrative domain, such as different EPC as shown in Figure 15 or different access network as shown in Figure 16, the UE should re-attach to the new wireless network, and obtain a new IP address.

It is obvious that the provider assigned IP address is always bundled with a specific access provider, a particular administrative domain, or a designated wireless access network. The IP address is only valid at an associated geographic area and for the duration of the control plane wireless attached with association. When a mobile device is moving to different location, the IP address of the UE will change, which leads to the IP based communication between mobile device and another device interrupted. This interruption can degrade the customer's user experience and reduce the service quality for service provider.

Figure 17 demonstrates this scenario when the IP address for UE is provider assigned address.

- Provider independent IP address

If a mobile UE uses a provider independent address, the IP address will be not change even if the device is moving to a different EPC or access network, but is constant for the PDN that provided the address. The problem caused by address changing is removed, but it introduces another problem that is the performance, stability and scalability for internet.

Provider independent address should be associated with a provider's network that the addressed device attached to. The BGP in the service provider's network should advertise the attached provider independent address to internet. Since the provider independent address block is not aggregate-able with the attached provider's address block; each end-user's Ipv6 address could be advertised into internet as a /128 prefix. This will dramatically increase the BGP table size. Moreover, if the device is moved to a new network, whole internet should be notified for this movement through BGP update. The complete population of BGP routes to whole internet is pretty slow. Considering more and more UE will be connected to internet, if any device attachment and movement may lead to the global BGP update, it is easy to imagine how bad the internet performance, stability and scalability will be.

Figure 18 demonstrates this scenario when the IP address for UE is provider independent address.

8.1.2.2 Scenarios for multi-homing and load balancing

When IP based addressing system is used for Multi-homing service and load balancing, the service is identified by one or multiple IP addresses. Similar to the above clause, the IP address can also be either service provider assigned or service provider independent.

No matter what type the IP address is, when the active link connecting to the service provider is down, the multi-homing IP address should be withdrawn from that service provider advertised routes, and re-advertised through another service provider. As a result, BGP withdraw and update for more specific routes (the multi-homing IP address) will be rippled over internet. Similar situation also applies to the case that multi-homing services are added or removed from internet.

This is one of the reasons that BGP routing table at DFZ (Default Free Zone) keeps growing in decades (IETF RFC 4984 [i.16]). It is a threat to the scalability of internet. With more and more multi-homing service added to internet, sooner or later the BGP table size may exceed some routers' physical TCAM size and the service provider should upgrade their network periodically. Figures 3 and 4 illustrate the models for this scenario.

8.1.3 Applicable Issues

Issue-01: **Non-Distinct ID and Address** In current networking and transmission protocol implementations, such as TCP/IP, the usage of IP addressing and ID has been overloaded for the following aspects:

i) 'addressing' = Network Communication Protocol Address, at each layer; and

ii) usage ID's: User-ID and Device-ID; and

iii) Session/Service-ID; and

iv) a full-application naming complement.

Recommendation: In a multi-access and multi-layer, context aware NGP environment future protocols should address the ID/Addressing aspects: Network Communications Protocol Address, usage ID's, Session/Service-ID's and Application Naming distinctly.

Issue-02: **Address/ID update complexity for Service and Session Continuity:** Today several mobility and multi-homing service updates include addressing updates which add complexity to the mobility, session management and service continuity aspects of the network, and also impact the scalability of internet especially the routing table in core routers.

Recommendation-01: NGP should minimize addressing updates in future protocols for mobility and multi-homing.

For this recommendation, it is noted that:

i) Application-ID should not change during mobility and multi-homing link state changes.

ii) Addressing may change but should be minimized.

Recommendation-02: NGP should aim to minimize NGP protocol complexity.

Recommendation-03: NGP addressing should support client-client, client-server (push and/or pull) and server-server models and multi-protocol versions thereof. This is in order to, for example: avoid multi-address-mappings to support NFV implementations, reduce addressing complexity to support Device-to-Device capabilities (D2D), etc.

For this issue, it is noted that:

i) There is currently no generic indirection in the mapping of addresses between well-known ports, the IP addresses, MAC addresses and other IP addresses. This makes it difficult to for example support mobility. This topic may be addressed in a future version of the present document or a derivative study, for example a section addition to NGP, WI3 or equivalent.

ii) NGP should minimize the use of "well-known" ports.

iii) It is assumed that NGP should include an addressing strategy that scales.

8.1.4 Applicable Use Cases

8.1.4.1 Case 1: UE communicates with a fixed device; UE is moving within a same P-GW domain

This is a mobility use case that a UE can communicate with a fixed device whilst the UE is moving within a P-GW domain. In this used case, the data service between the UE and the fixed device is not interrupted no matter how the UE is moving and where the UE is located. Figure 19 illustrates this use case.

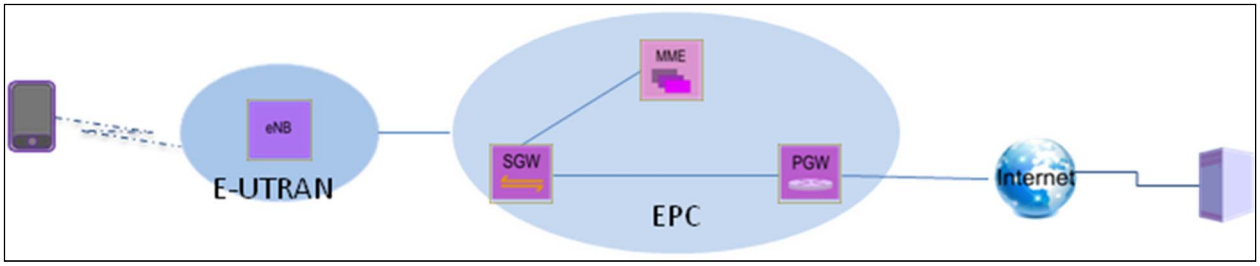


Figure 19: UE communicates with a fixed device; UE is moving within a P-GW domain

8.1.4.2 Case 2: UE communicates with a fixed device; UE is moving across different P-GW domain

This is a mobility use case that a UE can communicate with a fixed device whilst the UE is moving across different P-GW domain. In this used case, the data service between the UE and the fixed device is not interrupted no matter how the UE is moving and where the UE is located. Figure 20 illustrates this use case.

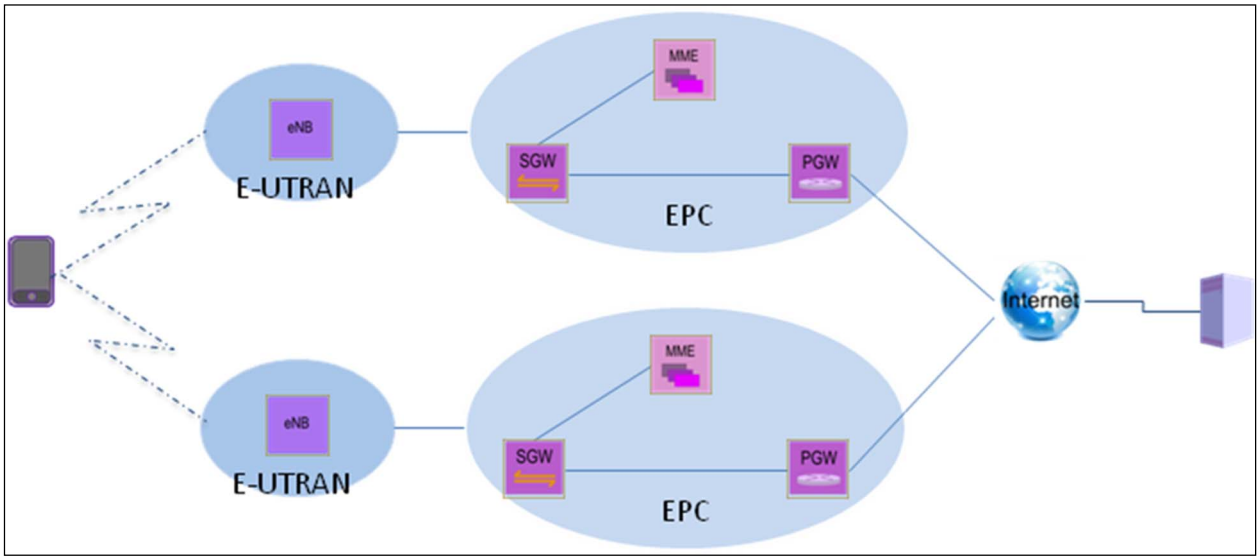


Figure 20: UE communicates with a fixed device; UE is moving across different P-GW domain

8.1.4.3 Case 3: UE communicates with a fixed device; UE is moving across heterogeneous access network

This is a mobility use case that a UE can communicate with a fixed device whilst the UE is moving across heterogeneous access network. In this used case, the data service between the UE and the fixed device is not interrupted no matter how the UE is moving and where the UE is located. Figure 21 illustrates this use case.

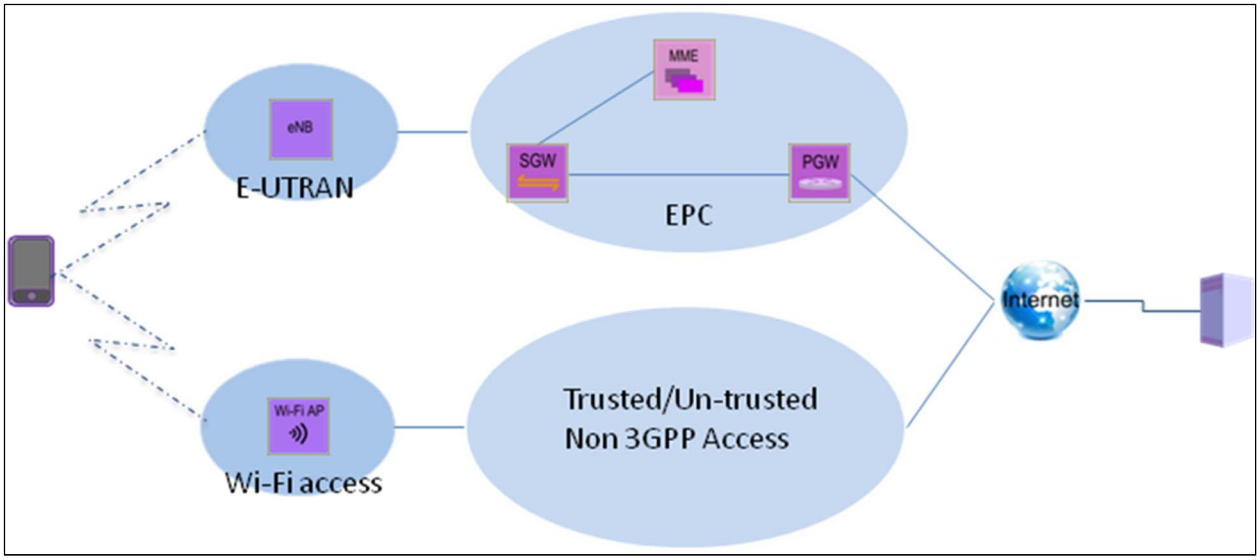


**Figure 21: UE communicates with a fixed device;
UE is moving across heterogeneous access network**

8.1.4.4 Case 4: UE communicates with another UE; UE is moving within a same P-GW domain

This is a mobility use case that a UE can communicate with another UE whilst the UE is moving within a same P-GW domain. In this used case, the data service between two UEs is not interrupted no matter how a UE is moving and where a UE is located. Figure 22 illustrates this use case.

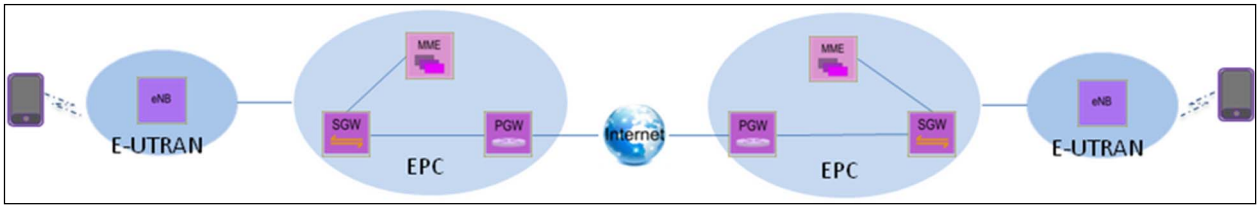


Figure 22: UE communicates with another UE; UE is moving within a same P-GW domain

8.1.4.5 Case 5: UE communicates with another UE; UE is moving across different P-GW domain

This is a mobility use case that a UE can communicate with another UE whilst the UE is moving across different P-GW domain. In this used case, the data service between two Ues is not interrupted no matter how a UE is moving and where a UE is located. Figure 23 illustrates this use case.

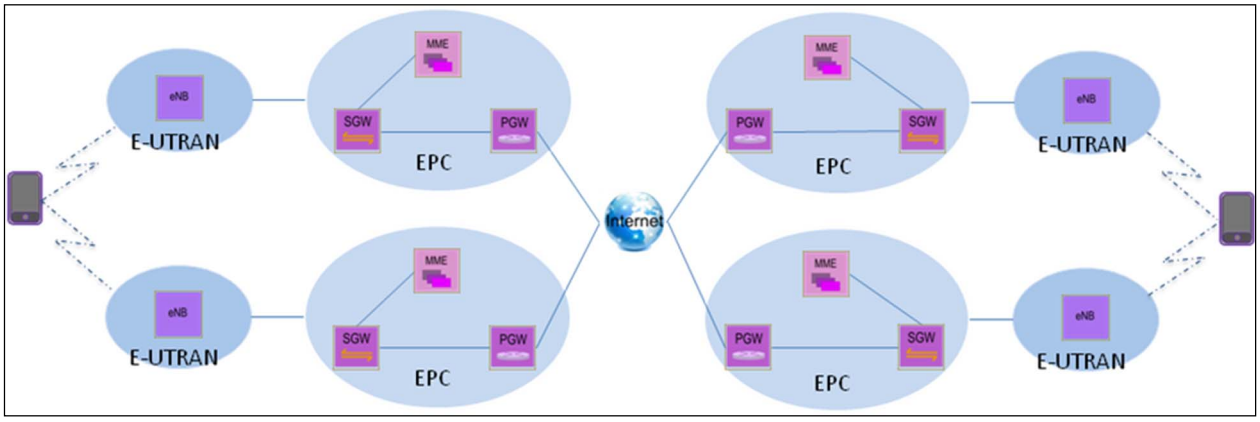


Figure 23: UE communicates with another UE; UE is moving across different P-GW domain

8.1.4.6 Case 6: UE communicates with another UE; UE is moving across heterogeneous access network

This is a mobility use case that a UE can communicate with another UE whilst the UE is moving across heterogeneous access network. In this used case, the data service between two Ues is not interrupted no matter how a UE is moving and where a UE is located. Figure 24 illustrates this use case.

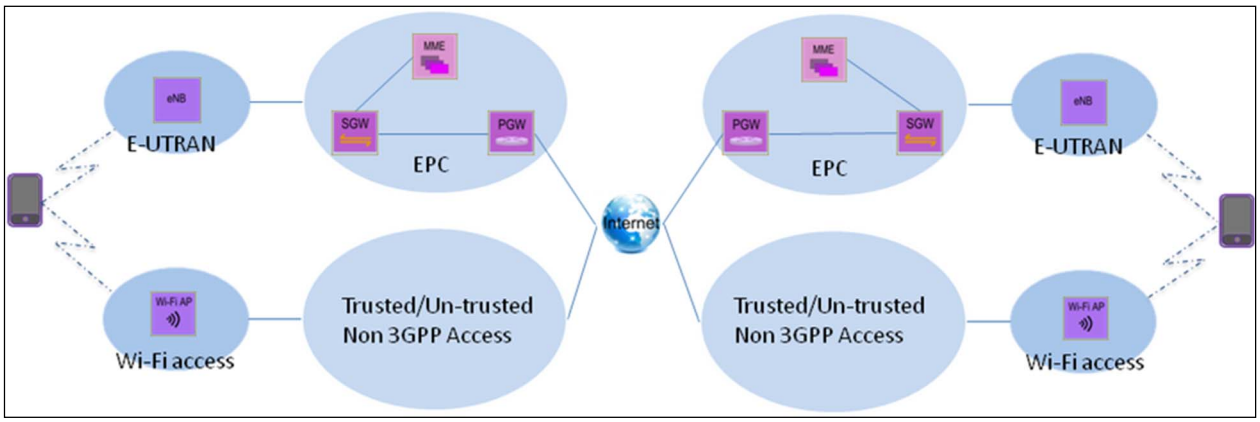


Figure 24: UE communicates with another UE; UE is moving across heterogeneous access network

8.1.4.7 Case 7: Multi-homing host connected to different ISP for link protection or load balance

This is a multi-homing use case that a multi-homing host connected to two ISP. The Host ID is used for the service provided by the multi-homing host. It does not change with the link status or the number of backup links. As a result, to use ID instead of IP as the identifier for the multi-homing service, the BGP routing table in the internet does not change with the number of multi-homing site, and the link status of the multi-homing site. Figure 25 illustrates this use case.

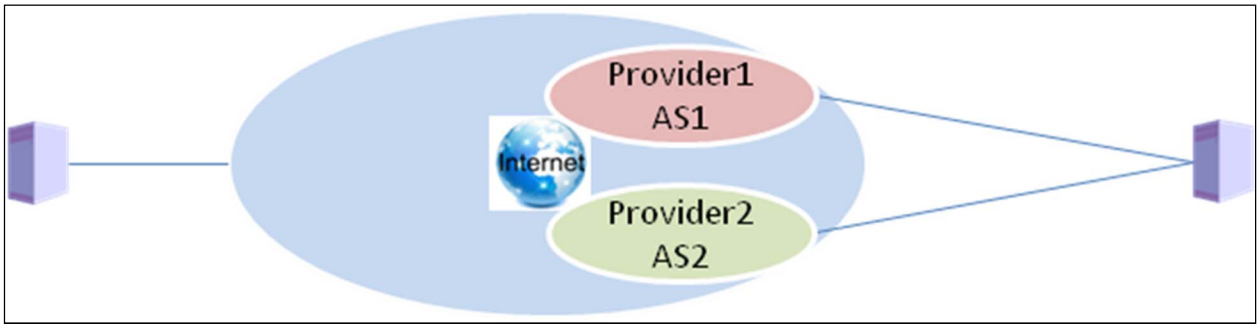


Figure 25: Multi-homing service with different ISP for protection or load balance

8.1.4.8 Case 8: Customer network with multi-homing site connected to different ISP for link protection or load balancing

This is a multi-homing use case that a customer network connected to two ISP, and the multi-homing site in inside of the customer network. The customer network address space is either assigned by one ISP or is provider independent. The Host ID is used for the service provided by the multi-homing site. It does not change with the link status or the number of backup links. As a result, to use ID instead of IP as the identifier for the multi-homing service, the BGP routing table in the internet does not change with the number of multi-homing site, and the link status of the multi-homing site. Figure 26 illustrates this use case.

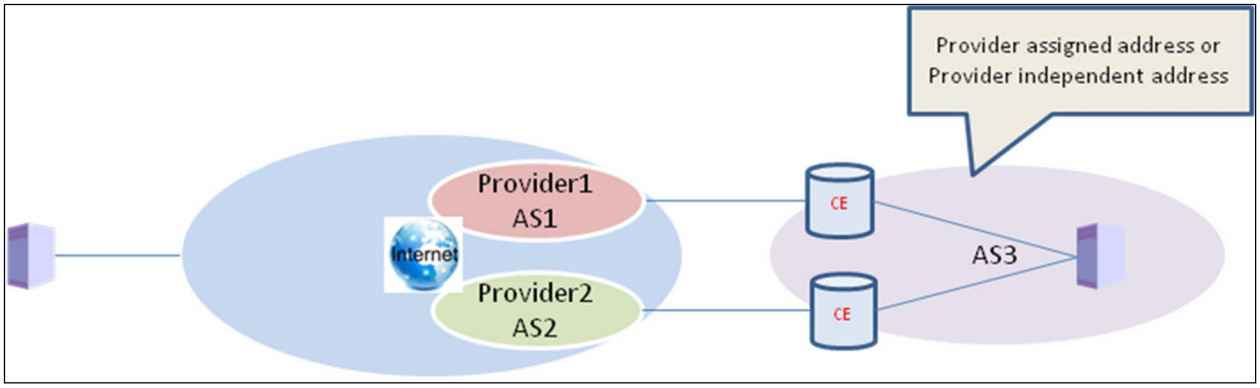


Figure 26: Multi-homing with different ISP for protection or load balance

8.1.5 Scenario Targets

Table 3 details the KPIs for improvement of this Scenario as a result of the development of NGP new addressing system.

Table 3: KPI's for Scenario - 01

KPI Name	Description	Measured feature	Units	Current Value	Target Value
Mob_fix_same_IP_ domain	Move UE within same IP address domain; peer is fixed server	Service	Binary Value	Service Interruption = 0;	Service Interruption = 0;
Mob_fix_diff_IP_ domain	Move UE across different IP address domain; peer is fixed server	Service; BGP routing table size at DFZ	Binary Value	Service Interruption = 1; BGP routing table size at DFZ change = 1	Service Interruption = 0; BGP routing table size change at DFZ = 0
Mob_fix_het_net	Move UE across heterogeneous network; peer is fixed server	Service; BGP routing table size at DFZ	Binary Value	Service Interruption = 1; BGP routing table size at DFZ change = 1	Service Interruption = 0; BGP routing table size at DFZ change = 0
Mob_mob_same_IP _domain	Move both Ues within same IP address domain	Service; BGP routing table size at DFZ	Binary Value	Service Interruption = 1; BGP routing table size at DFZ change = 1	Service Interruption = 0; BGP routing table size at DFZ change = 0
Mob_mob_diff_IP_ domain	Move both Ues across different IP address domain	Service; BGP routing table size at DFZ	Binary Value	Service Interruption = 1; BGP routing table size at DFZ change = 1	Service Interruption = 0; BGP routing table size at DFZ change = 0
Mob_mob_het_net	Move both Ues across heterogeneous network	Service; BGP routing table size at DFZ	Binary Value	Service Interruption = 1; BGP routing table size at DFZ change = 1	Service Interruption = 0; BGP routing table size at DFZ change = 0
MH_diff_ISP	Multi-homing with different ISP for link protection or load balance	BGP routing table size at DFZ	Binary Value	BGP routing table size at DFZ change = 1	BGP routing table size at DFZ change = 0

8.2 Security

8.2.1 Model Architecture/Protocol Stacks

The LTE cellular architecture illustrated in Figure 1 of the clause 6.1, shows existing Mobile Application Interfaces from 3GPP LTE Rel-12 which gives some examples of protocol stacks which are relevant to this security scenario (i.e. this is an example architecture rather than a formal model architecture).

Relevant examples from the referenced architecture include:

1) Ipsec for User Plane and Control Plane between Base Station and Core (Security Gateway).

2) Integrity message checking of all Non-Access Stratum (NAS) and Radio Resource Control.

3) Mutual authentication of User Equipment and Core Network over Non-Access Stratum tied to Authentication Centre.

4) Protected ID: IMSI only ever transferred as secured NAS tunnel using NAS Information procedure.

5) Access Stratum and Non-Access Stratum security updated every 5 keys from network.

Importance of network architecture in the security scenario:

For the most part, current security functions are tailor-made for each protocol (IP, TCP, BGP, DNS, etc.). When new protocols appear (e.g. LTE), new security functions are specified for this protocol. There are exceptions that point towards the path that should be followed (like EAP, the Extensible Authentication Protocol), which is to decouple data transfer and layer management (usually known as control plane) protocols from security functions.

This way the network designers could ideally pick from a/several catalogues of security functions (authentication, access control, encryption, integrity verification) and plug them where needed into its infrastructure. Of course, to do this right the network architecture needs to provide clear integration points for each type of security function - which highlights that network architecture is critical to efficiently and effectively secure a network.

8.2.2 Scenario Description

8.2.2.1 Scenario summary

Security requirements and expectations are changing. If new protocols are to be adopted, it will be driven by the business benefits perceived from 5G Use Cases. This implies that the security requirements should be supportive of 5G scenarios. The conclusion is that NGP security should satisfy three requirements. First, today's security challenges should be addressed in an increasingly efficient way to support new bandwidth and latency requirements. Also, new concerns and use cases should be accommodated without compromising core security principles. Last but not least, the cost of providing and managing security functions should be bounded, otherwise there may come a time in the future when the costs of being interconnected may be higher than its benefits (see [i.10] and [i.11]).

8.2.2.2 Security approach

The following principles underpin the security approach recommended for this scenario:

- Information is shared between layers only where there is explicitly a reason to do so. By default, information is not exposed between layers other than where it is demonstrably necessary. If a function is entitled to know information it should be relatively easy to extract (with the right credentials/audit) and if not it should be infeasible.

- The approach is based on security-by-design in that the intention is that the underlying design of the protocol should reduce the need for expensive (in terms of cost or network resource) security protocols and practices (specifically, removing the need for replication of security mechanisms). For example, a careful approach to addressing (see clause 8.1) can prevent address information from being shared more widely than is required.

8.2.2.3 Description of new security challenges

Clause 8.2.2.1 refers to new security concerns and use cases. These include:

1) Role in critical infrastructure. It is already the case that communications networks are being added to the list of components which are considered "Critical National Infrastructure" (CNI). However, this underestimates the role they will play. Networks will be vital to almost all of the existing components of CNI. Also, 5G networks will create and underpin entirely new components of CNI, such as Tactile Internet or Remote Monitors in health care and Vehicle Communications in transport. Availability and reliability requirements will be critical to the success of an NGP.

2) Privacy concerns. There is increasing public debate about the importance of user privacy and about who should or should not have access to user data. The debate asks which organizations (public or private) are trusted to hold user information and what they are entitled to do with it. The present document is not a philosophical or legal one and does not attempt to draw any conclusions from this debate. Instead the technical requirement is that NGP protocols shall support a range of outcomes from this public debate i.e. the technology should facilitate privacy where this is required and access to data where this is required. This implies that security may need to be handled differently depending on the component and information type in question (Personal, Network, Provider, Content, Government, Financial, Utility Control & V2x (…CNI), Personal-IoT).

3) Virtualisation and isolation. Many infrastructures running future protocols will be based on virtualised architectures. From a security point of view, the key consequence is that it will not be possible to rely on physical separation and therefore there should be an underlying assumption that data at rest and in transit will be visible to other actors (for example, hypervisors have access to memory of functions they are hosting; also network attacks may mean that there will be many compromised components running in the same environment as sensitive functions). Security-critical functions such as key negotiation or key storage will need to be built based on effective, strong isolation e.g. enforced through hardware roots of trust. As the network gets more virtual, a system that not only supports users secure access and privacy for their data but also formally logs all instantiations in terms of **What, When, Where, Why, and Who** is required so that it is possible to audit and trace issues in the first instance and as our information usage evolves, write security algorithms to monitor malicious behaviour. This is handled in more detail in clause 8.5.

4) Internet-Of-Things. These use cases impact security partly through scale: the number of devices to be authenticated will be an order of magnitude larger than at present. The devices will typically be built-in i.e. without human access (cars, meters, sensors) making it impractical to physically swap identity or security modules. Low-power IoT will have a significant impact: many traditional security techniques require considerable bandwidth (e.g. for handshaking even if not for traffic delivery). Also, connected IoT devices potentially provide a way of bypassing security measures (such as firewalls) protecting other equipment on the same network.

5) Network Optimization. Where security components have been "bolted-on" on top of existing protocols, there can often be inefficiencies with additional proxies and protocol layers removed and re-applied. Such inefficiencies could defeat the core benefits (bandwidth/energy/manageability/reduced complexity) which form the heart of what NGP is trying to achieve. Network optimization is only possible where carriers are able to understand key meta-data from the traffic they are conveying. A carefully thought-through approach to confidentiality at an enterprise level can enable operators to see the information they need without exposing excess information. This is handled in more detail in clause 8.7.

8.2.3 Applicable Issues

Issue-01: **NFV:** As networks and particularly Cellular Access networks migrate from implementations of Network Elements (NE) built on bespoke hardware platforms to Virtualised Network Functions (VNF) realized on Off-The Shelf (OTS) computing equipment as is likely during the 4G to 5G era, then the infrastructure becomes more flexible, easy to evolve but also more susceptible to security threats. ETSI NFV ISG has reported on the NFV/SDN security aspects in their reports highlighting how the infrastructure is less secure than it used to be as it is operating over much larger depth and breadth without the existing hardware boundaries in place today; from a much wider potential scope of control on hardware that is likely to be provided by a non-operator tenant in many cases and in some cases by a multi-tenanted hardware provider.

Fundamentally, the challenge is that virtualisation removes existing hardware boundaries between sensitive functions. This leaves "all eggs in one basket". Consequences include:

- Successful cyber-attacks may more easily gain access and control over the entirety of a network (including password or cryptographic stores, sensitive personal data).
- It is not possible to keep sensitive stores of information isolated from the rest of the network. For example, anyone with hypervisor privileges could access all sensitive personal/commercial data. See (ETSI GS NFV-SEC 009 [23]).

In general, these are not totally new or unstudied problems. There are some similar security considerations as in the "multi-tenant" cloud security problems. The difference is extent of including Core Network and RAN Infrastructure.

Recommendation-01: The new security capabilities that are needed to support virtualised infrastructure in an NGP context are as follows:

- Efficient extensions for MANO and SDN controller protocols should be provided to administer them across Tenant and Multi-Tenant environments in a secure manner (mandatory two-way authentication by traceable ID, SW event logging and resource management monitoring at the memory, processing, VM, Flow and VNF levels. It is critical that these are built-in controls that are mandatory rather than optional bolt-on controls.
- Confidence in components should be tied to the use of hardware root-of-trust attestation.
- Separation of sensitive components: an architecture with separate trust domains for key sensitive functions should be incorporated. Ensure that, for example, fraud management, authentication/crypto credentials, cyber defence, law enforcement functions (Lawful Intercept) are all managed independently and that access to one does not grant access to all information (ETSI GS NFV-SEC 009 [23]).
- When NFV incorporates open source software; there should be procedures in place to ensure that basic security practices keep pace and are reflected in open source software.

This issue is linked to security for ID, Authentication, monitoring and configuration (see later issues).

Issue-02: **MEC:** From a security point of view, there are many similar challenges from Mobile Edge Computing as there are from NFV. The central issue is that more functionality will be managed from environments with much lower levels of physical or hardware security. The issues are therefore similar to those from NFV, see Issue-01 (for example, MEC architectures will need to be designed to meet Lawful Interception requirements).

Recommendation-01: With the introduction of MEC , attestation of components and separation of sensitive functionality should be observed to maximize robust security. (i.e. again, this is similar to the recommendations in NFV - see Issue-01).

Recommendation-02: MEC generates requirements for operators to have a clear view of communications meta-data, in order to route traffic efficiently. As per Issue-07, it is important that MEC meta-data should be handled securely with access when required and only when required.

Issue-03: **Energy efficiency:** There is a strong trend towards use of low-power sensors as part of the Internet-Of-Things. In some scenarios, sensors are inaccessible and small and so batteries should be small but should last many years or decades. This can place very tight requirements on bandwidth, memory and CPU which means that modern security protocols are not appropriate (e.g. those with high-bandwidth handshaking procedures or the ones that require CPU-intensive cryptographic operations). For mainstream situations there is naturally still a focus on energy efficiency in terms of cost savings.

Recommendation: Internet-of-Things security designs/architectures should be aware that significant components of their networks may be on very low power which would affect the bandwidth, memory and CPU that could be allocated to security protocols. It is not the case that all IoT security should be run at low bandwidth, so any NGP security protocol suite would need to contain both high-resource-consumption and low-resource-consumption protocols.

Issue-04: **Identity, Authorization and Authentication:** There is a public debate about the extent to which a personal identity should be assured or confirmed prior to providing a service. Services which randomize or rotate identities will add to user privacy; they may reduce efficiency, they are likely to compromise audit/accounting functions (see below) and will remove compliance with many regulatory regimes.

The specific issue around NAT: Network Address Translation functionality was introduced (in general) due to pressures on the Ipv4 address space. A consequence of NAT is that there is some obscuring of internal identifiers from public traffic. Also, there may be a case that operators gain benefit from NAT due to the way it can obscure network architectures (beneficial for commercial or security reasons, potentially). Any next generation protocol would need to be designed so that issues with running out of addresses are avoided. There needs to be a debate about whether NAT-type functionality would be desirable or cost-effective as technology evolves to a solution where there is no pressure on the address space.

NOTE 1: The adoption of a complete addressing model as recommended for 8.1 will eliminate the problems caused by NATs as noted in this clause without eliminating any of the advantages on NATs. (NATs only break broken architectures.

Recommendation-01: Adding an Identity Service (logging/checking/validation, management) would be of benefit in avoiding some security vulnerabilities in the current IP suite. Consideration may need to be given to ID servers and/or Resource Registration (RR) for logical entities (see also IdoT and IdoP). Examples include ID servers for entities such as VNFs from the NFV space, or a register of Certified IoT devices or IoT Gateways (see next paragraph). Consideration should be given to techniques for providing authentication or attestation without the expense and complexity of running a full PKI.

Recommendation-02: The role of IoT Gateways should be investigated as part of providing secure Identity services in NGP. As latency requirements get tougher, the pressure for edge-based decision-making will increase. Gateways can have an important role in facilitating secure edge-based computing, potentially enabling better security and/or more effective use of meta-data (see clause 8.8).

Recommendation-03: The ID system should to add security but not be prohibitively expensive. There is a balance of security versus cost and complexity to be considered in satisfying this recommendation, so it is recommended that NGP evolves to scalable security and ID system where cheap and large scale edge devices have some level of ID checking and the gateways or concentrator nodes that they feed into have a higher level of ID security, typically RADIUS based for example. NGP needs to consider scalable ID capabilities and recommend failsafe mechanisms higher in the network to manage the threat level.

Recommendation-04: NGP should recognize that compartmentalization will be required for certain kinds of endeavours, especially with the limitations of IoT. The NGP model should be able to utilize isolation.

Issue-05: **Location:** Location is at the heart of many of the benefits which 5G aims to deliver. Applications which use location to generate revenue or deliver new user experiences (proximity, direction finding) are already very important and increasing due to trends such as vehicle communications. Location as an essential part of regulatory concerns. For example, financial or legal institutions may have a requirement to understand where contracts or transactions took place. It can be very important to know and share the location of an emergency call without delays.

Specific issue around accuracy and reliability: There are examples today which highlight some of the concerns around location accuracy. GPS is normally 50 m accurate and then is enhanced by a bootstrapping function (DAB, B3, Google-AP-Ref, etc.). Indoors, the state-of-the-art is 2 m accuracy within 100 m reference using Mobile Sensors. It should be noted that these functions are not guaranteed to be available and some of them (e.g. AP referencing) can be prone to mis-use.

Recommendation: There should be a clear assessment of the situations where location meta-data needs to be propagated and where it should remain private or be obscured. Network protocols need to support situations where location meta-data is a part of network audit or where it is required to be delivered for business purposes. It should be noted that location data can contain considerable personal or sensitive information and location information should not be transferred or made available without a proper assessment of the privacy implications.

Issue-06: **Accounting and Audit:** It is clear that per-use billing is decreasing rapidly but it would not be correct to imply that there is a reducing need for accounting or audit functions. Key business drivers for accounting/audit functions include fraud management, monitoring of users' compliance with their contracts and audit of business-to-business relationships. Compliance with national regulations (such as assistance to Law Enforcement) requires clear audit functionality. Accounting and Audit information should also be able to be checked by the user where appropriate. The solution needs to be scalable, traceable and cost effective.

Recommendation: Security assessments and standards often include references to security monitoring (frequent checking on security-related/impacting functions). Security monitoring should be an in-built part of NGP.

Issue-07: **Meta-data and APIs:** This issue looks at the business needs for visibility of traffic meta-data. It looks at the requirements for knowledge at a given layer of the protocol stack for meta-data generated by other layers.

In current networks, there are applications in which traffic has to be inspected at a given layer in order to extract parameters from other layers (e.g. network monitoring, traffic shaping, DoS prevention). This implies that the current "network service API" is broken. Ideally NGP should evolve to a stage where it is not necessary to inspect traffic but rather request necessary parameters from the control or management functions. Today's approach (around traffic inspection) can make it difficult to meet legitimate requests or needs for meta-data but also facilitates access to meta-data which is not authorized/audited or known about.

The issue in current networks, using the example of optimization information is that currently options for optimizing an IP network and/or a mobile network involve inspecting traffic or requesting information over interfaces which are not designed for this purpose. Examples include:

a) Updates to the Operations/Management Centre (very slow e.g. every 15 minutes).

b) Inspection of mobile/client user information (note there is no standard structure today).

c) Run traces at devices and nodes (impacts performance on the entity and is bulky and slow - line speed not filtered).

d) Interwork with bespoke/proprietary IoS (slow).

e) Examine SNMP interfaces (usually they are proprietary with obscure Ims).

f) Use the BBF TR-069 [i.27] interface for set-top boxes (slow).

g) Operate SSH (very slow and inhibits most features needed for management, control and orchestration).

Recommendation-01: Each layer should offer an API that allows the layer above to request the properties of the network service it wants (bound on packet loss and delay, in-order delivery of data, etc.).

Recommendation-02: Layers should have good layer management and admin that provides statistics needed for security. It would be important to use different Information Object Classes (IOCs) for different protocols: NGP needs a control structure that enables a layer to select which IOCs can be accessed dependent upon context of User, Network and Target peer (from no IOCs to many IOCs) on a per communication binding basis.

Recommendation-03: Follow-up on optimization: In contrast to the current options listed above, optimization Meta-Data could be based on a "pull" meta-data system. However, pull is usually slow (to avoid slowing down the user plane paths). Instead optimization could be provided as a separate Meta-Data mechanism for which entities can offer information to, "push" meta-data. This is a mechanism that is specifically designed for shipping context information around for an operator or user to optimize their user QoE or Network performance. Also see: Virtualisation Scenario.

Recommendation-04: A more specific requirement for meta-data APIs is likely to come via the concept of "network slicing". Next-generation solutions will be operating over heterogeneous networks and with varying network requirements. This creates a drive towards building network functions as slices, created out of a number of sub-components or sub-functions. Coordination of meta-data across these virtualised components should be supported by NGP to meet the goal of information being available where needed and only where needed.

Examples include:

i) Flow allocation: in which the layer below gets a flow request (a flow is an instance of a communication service from the layer above (which may be an application). The request contains a set of meta-data describing the requirements for the flow (bounds on loss and delay, in-order delivery of data, etc.). This information is used by the layer allocating the flow to select the best configuration for the resources that the flow will use, according to the resource allocation policies available in the layer.

ii) Cell boundary management: A protocol to show where the traffic is and then provide to an operator to steer cell boundaries to load balance. This is an example of a hybrid self-organizing network system.

Issue-08: Applications need to protect the confidentiality and integrity of their communication.

Recommendation-01: Applications should protect the confidentiality and integrity of their communication.

Recommendation-02: NGP recommends that each layer and/or the protocols in each layer should take the necessary security measures to protect their data without relying on the lower layer.

NOTE 2: Most security experts agree that the best encryption architecture should operate between peer applications.

NOTE 3: Operating peer application encryption greatly reduces the security problem for the network to primarily authenticating members of some layers, and protecting against traffic analysis at higher communication layers.

Issue-09: **Data Transfer Protocols:** Port and Connection Endpoints in the current IETF: TCP/IP protocol definitions are combined.

Recommendation: NGP should adopt data transfer protocols that decouple port-id from connection-endpoint-id and avoid so-called 3 way handshake synchronization sequences and avoid well-known-ports.

Issue-10: **Securable Containerisation** is not currently used.

Recommendation: NGP should adopt structures, such that an application and its correspondent form a securable container. Similarly, as required all layers should be securable containers.

Issue-11: **Trusted Hardware Bases and Trusted Local Resource Allocation** are not currently widely used and the current form of open source software cannot be secured.

Recommendation: NGP providers should move toward trusted hardware bases and trusted local resource allocation, e.g. operating systems.

NOTE 4: Securing the network without securing the systems that the network software runs on is a waste of investment in the network.

A further set of examples is listed as part of the KPI's in clause 8.2.5.

8.2.4 Applicable Use Cases

The issues listed in clause 8.2.3 should be assessed particularly in the context of the following Use Cases: see Annex A.

- 5, 6, 7 and 10 regarding mobile broadband.
- 4, 16, 23, 48 and 51 Co-existence with legacy work is going to be important.
- 27 and 28: Multi-access/multi-RAT will have an impact.
- 32 and 33: Vehicle situations. It's worth thinking about whether the vehicle situations (32 and 33) have any specific security required.

It is noted that in general greater connectivity or bandwidth is less of a concern from a security point of view, (though some considerations do apply e.g. see Issue-03 of the Security Scenarios).

8.2.5 Scenario Targets

Table 4 details the KPIs proposed for improvement of this Scenario as a result of the development of the ETSI ISG NGP.

Table 4: KPI's for Scenario 02

KPI Name	Description	Units	Current Min Value	Current Max Value	Target Min Value	Target Max Value
KPI-01: Latency for security protocols	Assess impact of security protocols on latency requirements.	Should be considered in line with existing KPIs for latency.				
KPI-02: Generic security cost assessment	Business cases should provide justification for cost of implementation of security protocols and features.	£	Should be assessed for each business or business sector (e.g. see framework below).		Should be assessed for each business or business sector (e.g. see framework below).	
KPI-03: Meta data availability	See clause 8.5 and Issue-07 of this scenario for description of meta data issue. First requirement is that meta-data should be present; a further requirement is the efficiency of providing this data.	Measure in seconds delay *and* cost to operator of failing to have this info.	See example in clause 8.5 and Issue-07 of this scenario - examples include 15 minute delays.		See target scenarios to be efficiently supported in notes below this table.	

Detailed notes regarding KPI-02:

The following text provides some notes on a framework for assessing cost/impact of security. There are many more formal approaches to assessing risks and, for each particular business sector or business, existing detailed frameworks should be sought out and used (e.g. note the 3GPP LTE requirements document for security architectures following a threat assessment for each facet of security).

Consideration should be given to the number of users affected (single user up to entire EPC or operator network) and the cost per user. This would cover aspects such as:

- Mutual user and communications peer ID (whoever authenticates the ID of the peer App, service or user) Authentication & Authorization
- Mutual user Equipment (UE) and communications peer equipment (CPE) Authentication & Authorization
- Encryption of stream or flow
- Operations and Management (AM, FM, CM, PM, SM) for equipment and software

For aspects such as Data or Message Integrity, it would be important to consider how long such threats could go undetected, and how many users would be affected.

Detailed notes for KPI-03:

The following scenarios need to be efficiently supported by the availability of the meta-data listed:

1) Network monitoring and optimization (traffic shaping). See Issue-07 of the Security Scenarios, paragraph on "optimization example" for details.

2) Security monitoring. Network monitoring to detect cyber-attack. Also, for issues such as parental control.

3) Maintenance of Critical National Infrastructure. The ability to identify critical traffic and to identify rapidly unwanted/malicious traffic e.g. DdoS attacks.

4) User device information. Security devices such as enterprise firewalls can make better or quicker decisions with knowledge of device meta-data (type of device). Meta data about user devices would facilitate or accelerate discovery/handshaking negotiations; this may be critical in low-power Internet-of-Things situations.

5) Digital Rights Management and protecting Intellectual Property.

6) Use of meta data for legitimate commercial and business purposes. Many communications companies gain business benefit by understanding more about their customers' behaviour. There is a commercial pressure to enable this, to the extent permitted by privacy regulation and in line with agreements made with end users.

7) Regulatory requirements such as Lawful Interception and Data Retention require access to meta-data, to be followed in accordance with national regulation.

8) Location information (see Issue-05 of the Security Scenarios) and mobility information such as speed will be important from a network management point of view, particularly for high-speed 5G Use Cases.

8.3 Mobility

8.3.1 Model Architecture

This scenario is described so as to identify the issues that should be considered for current and evolving user access networks with respect to next generation protocols, that provide communication for users towards the Internet and other PDNs where there is a need for scalable, any-access mobility to be supported.

The scenario is illustrated using key bindings that should be accommodated by the NGP ISG when considering this scenario.

This scenario description firstly introduces the basic support model that should be provided for a single current 3GPP, LTE-A Release 12 mobile radio access network in order to support scalable mobility. This access network is illustrated in terms of the key bindings that should be provided in Figure 27.

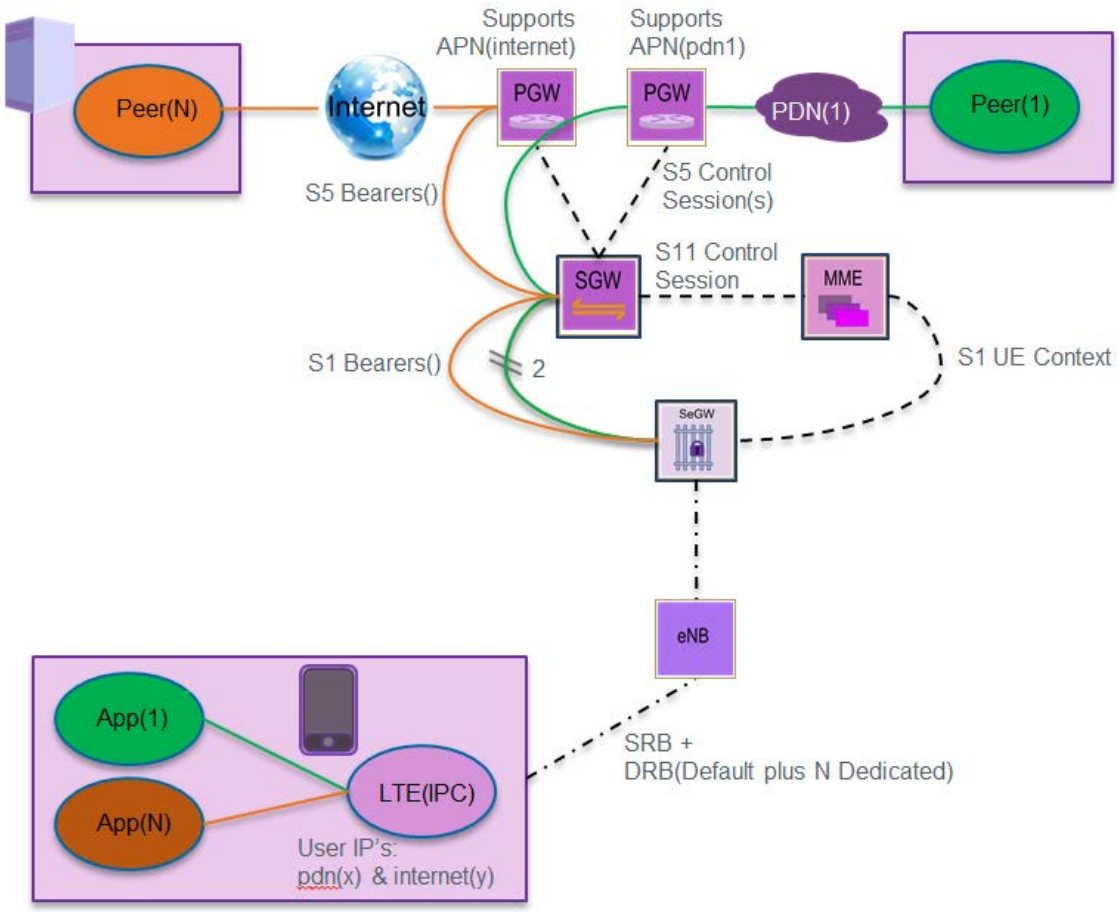


Figure 27: Release 12 LTE Access Network Support

Figure 28 introduces an evolved model that should be provided to support a single 3PP, LTE-A Release 14 mobile radio access network in order to support scalable mobility, where provision for access on a Dual Connectivity basis is required.

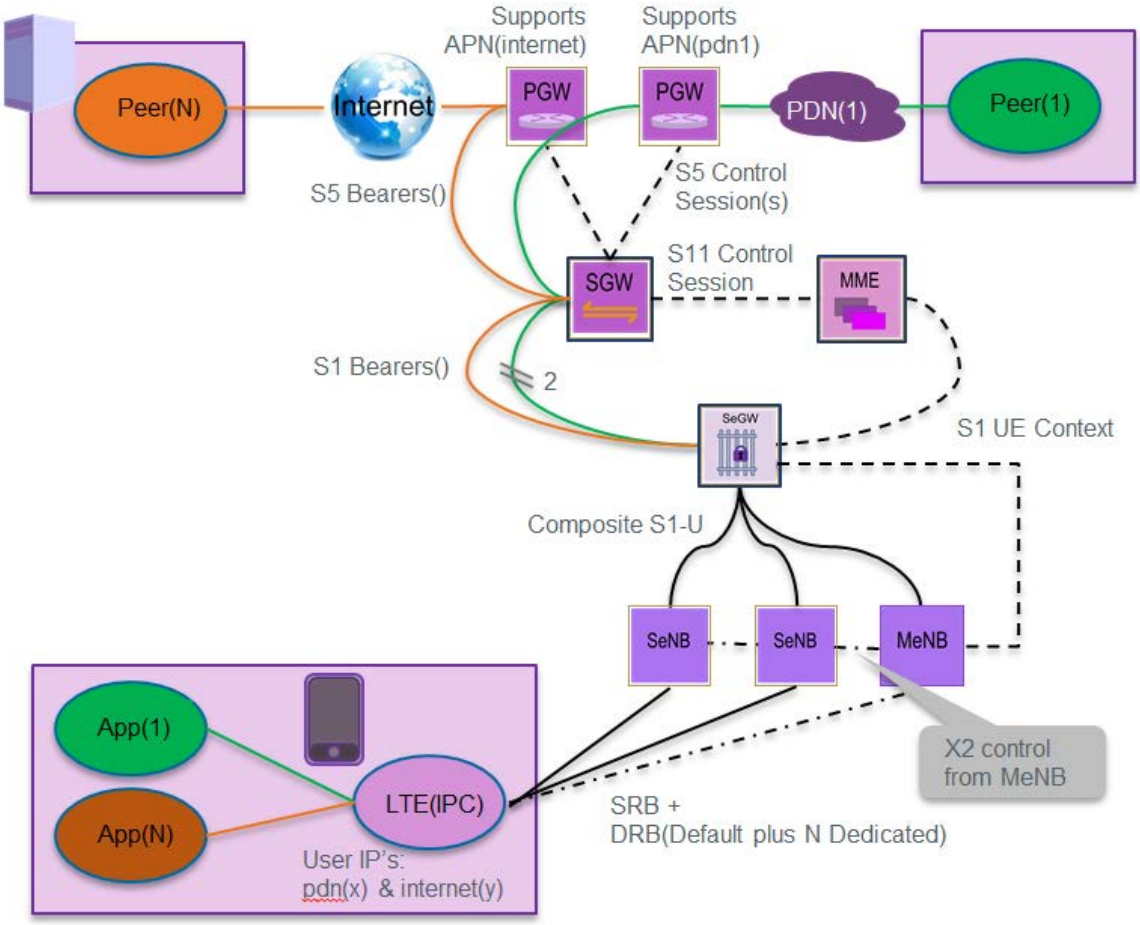


Figure 28: Release 14 LTE Access Network Support, with Dual Connectivity Support

Figure 29 introduces an evolved model that should be provided to support multiple concurrent access networks at the same time in order to support scalable mobility and where the CP and UP services provided to the user may not be provided by the same access network.

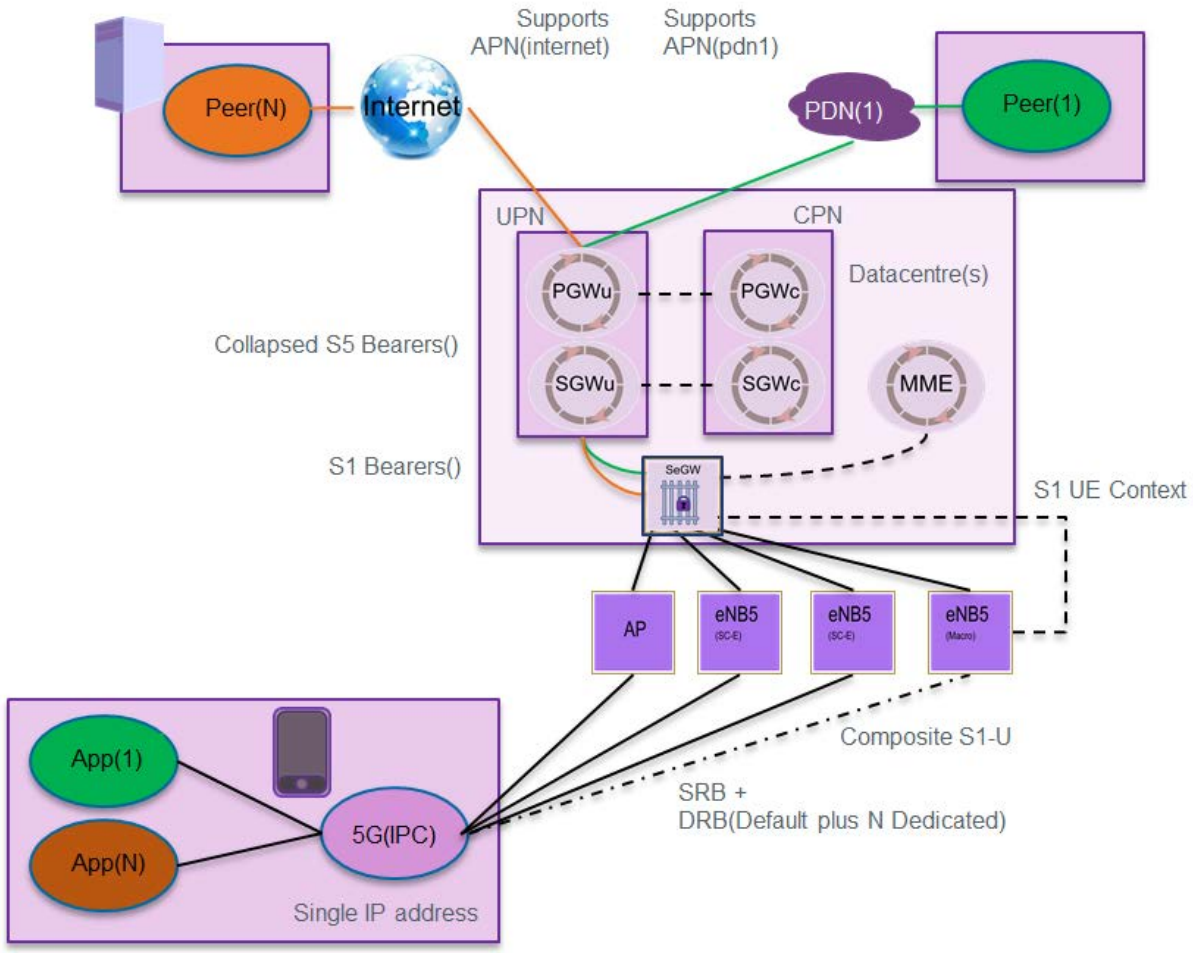


Figure 29: Multiple Concurrent Access Network Support with Fully Separated UP & CP Scalability

For this evolution of the multi-access network the user plane entities are all col-located in a logical node called the UPN and the control plane entities are co-located in a logical node called the CPN. The CPN to UPN core interfaces are new (yet to be defined in the 3GPP Release 15 CUPS work item).

In this scenario evolution multiple APNs can be supported whilst the user only has one IP address on the operator network.

There are several advantages in this evolution:

i) The S11 Control session is local (less latency).

ii) S5 Control session is collapsed (less latency).

iii) Multiple access links are able to be operated for the user concurrently across multiple supporting access technologies to provide a compound access connection which is the sum of the access links.

iv) The Control, and user plane are totally separated over the access technology and may even be on different technologies (enables CP and UP independent scalability and ability to select the best Access type for CP and UP).

v) The User is able to access all of the envisaged access technologies in 5G such as Wi-Fi™, Cellular RF, Cellular mm Wave and Fixed.

8.3.2 Scenario Description

This clause further describes the support required for the models in this scenario.

Mobility should be provided whilst supporting multiple applications on an ETE basis and handling multiple QoS sessions at the same time.

For the purposes of the present document, Mobility shall include standard 3GPP definitions of "camping" and "user mobility".

Mobility shall be provided so that a user device may connect to one or more access networks of an N-concurrent access network eco-system available to them. For the purposes of the present document camping is defined as being able to setup a logical connection between the user device and one of the N-concurrent access networks.

Mobility shall be provided so that whilst a user device is connected to one or more access networks of an N-concurrent access network eco-system the user device may move geographically relative to the fixed infrastructure providing N-concurrent access. For the purposes of the present document this behaviour is known as user mobility.

During periods of user mobility, the network shall be able to respond to changes in logical connectivity for any of the access networks currently in use without a reduction in service continuity, reduction in QoS for any of the supported ongoing applications supported over the N-concurrent access network eco-system or CP service outages.

Changes of logical connectivity between the User and physical access points to any of the access networks involved in the current composite access connection are collectively known as "handover".

During periods of user mobility there should be a maximum UP service gap of no more than <UP-Gap> milli-seconds and a maximum CP loss of service of no more than <CP-Gap> milli-seconds.

A user may be connected to up to <ConcurrentAccessConnections> x N-concurrent access networks at any one time.

A user should be able to handover <ConcurrentAccessHandovers> at a time without exceeding the maximum allowed UP or CP service gap specified.

Where: <ConcurrentAccessHandovers> is less than <ConcurrentAccessConnections>.

For the LTE, Release 12 model in Figure 27, a handover involves one logical connection with one eNB.

For the LTE, Release 14 model in Figure 28, supporting dual connectivity, a handover involves:

i) one logical handover of the RRC connection with the Master eNB (MeNB) (source) to MeNB(Target);

ii) the teardown of $\leq$ P x X2 links to all of the other S-NBs of the SeNB group; and

iii) the re-establishment of $\leq$ Q x X2 links to all the new S-NBs of the new SeNB group.

Where:

P is the maximum number of S-eNB in the source S-eNB group;

Q is the number of new S-eNB links established for the target S-eNB group.

For the Multiple Concurrent Access Network model in Figure 29. A handover may involve:

i) one logical handover of the access (e.g. RRC for LTE RF cellular) connection to the currently nominated access technology supporting the CP from source access point group on the access technology to target access point group on the target access technology; and/or

ii) one or more logical handovers of UP access connection(s) for any of the access technologies that need to be handed over as a result of the users mobility from source access point group on the access technology to target access point group on the target access technology.

8.3.3 Applicable Issues

Issue-01: **User Identity:** The user identity is often not preserved in communications networks necessitating multiple naming conventions to be stored and translated.

Recommendation: The user shall be able to retain the same identity as presented to the application layer during any mobility operations.

Issue-02: **Device Identity** is often not preserved in multi-access communications networks necessitating multiple naming conventions to be stored and translated for a compound handover.

Recommendation: The device shall be able to retain the same identity as presented to an N-Concurrent Access communications network during mobility operations.

Issue-03: **Communications Address:** The communications interface at each access point shall have a separate communications address to its name and location, often as in IP these are not separated which causes issues with mobility.

Recommendation: Each access point shall be able to be identified by its address, location and name.

The handover control mechanisms shall be able to identify the available access technologies, access points providing access to each technology and logical naming structure that groups access points together such as the terms of Cell and Location or Routing or Tracking area, for cellular technology.

Issue-04: **QoS classes and Traffic Flow templates** have been defined for many access technologies, but these are not often exposed to the user and are invariably over complex to handover as a compound set of bearers.

Recommendation: Multiple bearers should be able to be supported at the same time over the same compound connection whilst addressing different external PDNs during a handover. Bearer QoS and TFT should be exposed to the user or at minimum easily accessible to application developers and be able to be either setup explicitly by the user or dynamically according to subscribed or user defined profile during handover.

Issue-05: **Tunnelling** for the sake of mobility should be avoided wherever possible, as tunnels add notable overheads to each packet and incur Layer 3 and 4 operations to be able to move to execute a handover.

Recommendation: The NGP ISG should consider the best combination of one or more of the possible methods for providing mobility:

i) translation;

ii) tunnelling; or

iii) routing Table update;

in order to improve mobility for next generation systems.

The ISG should carefully consider the strengths and weaknesses of each mobility technique listed above, before developing a new technique.

Issue-06: **Scalable Mobility:** Any NGP evolution for mobility should be scalable, so as to support devices that are static as well as devices that operate at speeds increasing through walking speed, car speed to High Speed Trains (HST) over the same N-concurrent access network eco-system. Currently 3GPP networks support handover using RF dominant information to drive mobility and do not harvest the other information available on the device to be input to the mobility decisions of the communication network except the recent special case of a Mobile Category specifically for Static IoT devices which has just been defined in Release 13. Also, there is no mobility context that crosses access domains, for instance between Wi-Fi™ and Cellular.

Recommendation: the NGP ISG should consider a context aware mobility solution that can respond to the current and evolving mobility context of the device in terms of: location, speed and heading in the setup of any compound connection towards the N-concurrent access network eco-system and its subsequent mobility thereafter for the duration of the compound connection.

Issue-07: **Mobility Scope:** Currently the Scope of mobility is well defined in 3GPP using the Cell, x-Area (LAC, RAC, TAC) and Operator Network (e.g. PLMN, Wi-Fi™ or Fixed access network). However, with the advent of Heterogeneous Cell deployments: Macro, Micro, Pico. Femto and multi-point cells even this well-structured approach does not cover the evolving scope control of mobility.

Mobility to the accuracy of Location (to a few Metres) could be seen as an excessive mobility pointer reference accuracy for identifying a device to access point mobility binding, however some indication of proximity to cell centre of at least which AP of a multi-point cell should be considered for future evolutions of NGP.

With the advent of LTE indication of both Cell Centre and Cell Edge was included. However, even this capability gives no 2D indication of where in that Cell Centre or Cell edge the user is, e.g. Cell Centre, North West quadrant.

Recommendation: In order to make mobility efficient, some further evolution of mobility scope that adds "quantized location in cell" and "AP(x) of cell" to the existing common Cell, Cell Cluster or Operator Network established mobility hierarchy needs to be considered.

Also, in order to operate multi-access mobility as a scalable solution there should to be a common/easily translatable notation across access technologies.

8.3.4 Applicable Use Cases

8.3.4.0 Introduction

The following mobility specific use cases are developed from the SMARTER referenced use cases in Annex A. SMARTER referenced use cases: 5, 6, 7, 10, 12, 16, 28, 29, 30, 32, 34, 52 and 47 are identified as having notable mobility scenario dependence and/or are required in order to support next generation mobility enhancements.

8.3.4.1 Case 1: Multi-Access, Session & Bearer connection, Same Macro

A user device is compound connected to an N-Concurrent Multi-Access Network with concurrent links towards the best serving Macro, RF-Cellular Cell for CP coverage and operates UP bearers towards the Macro cell for wide area basic UP coverage and service continuity between UP during handovers between local supplemental UP providing APs. The user device also has localized supplemental user plane coverage provided through an access connection to two other cellular small-cells and has a supplemental UP connection towards a local Wi-Fi™ coverage access point.

The user device has several UP bearers in operation across the composite connection it is operating towards the N-Concurrent Multi-Access Network which in turn support several different user sessions, which include intermittent video streaming, background social networking and email, frequent web page browsing and at this time there is an ongoing IMS call in progress.

The User is moving at walking speed and moves from good coverage to all of the original cells through an area of coverage where they are only within coverage of the same Macro and have moved to different Small cell coverage and there is no Wi-Fi™ coverage then back to a position where they are within coverage of both small cells, and Wi-Fi™ again within the space of 5 minutes.

Successful operation means that the IMS call is held up for the duration of the 5 minutes use case and the bandwidth of the other sessions is managed s that there is no CP service interruption gap, SI(CP) and no UP service interruption gap of more than SI(UP) for any of the ongoing 5 services.

The average handover failure rate performance for multiple users operating the same kind of scenario is less than HI_Rf for more than 1 000 similar use case executions.

This use case includes the following service and network referenced use cases from Annex A: 4, 5, 6, 7, 8, 9, 10, 22, 26, 27, 28, 30, 35 and 47.

Note this use case is equally applicable if the Wi-Fi™ supplementary UP access service described here is either a Wi-Fi™ or mm-Wave access technology, in practice.

8.3.4.2 Case 2: Multi-Access, Session & Bearer connection, with Macro HO

As for case 1, but where coverage also changes from Macro(a) to Macro(b) for support of the CP connection to this Macro and its wide area coverage UP service whilst the user is moving.

8.3.4.3 Case 3: Single Access, Session & Bearer, Same Macro

As use case 1, but for only one mobile access technology provided by Cellular RF Macro and Small Cell coverage, where the user device is operating either a video streaming, background social networking, email or web browsing session.

This use case includes the following service and network referenced use cases from Annex A: (all from case 1).

8.3.4.4 Case 4: Single Access, Multi-Session, Multi-Bearer, Same Macro

As use case 3, but with multiple sessions and bearers.

8.3.4.5 Case 5: Fast, Single Access, Multi-Session, Multi-Bearer, with Macro HO

As use case 2, but for only one mobile access technology provided by Cellular RF Macro and Small Cell coverage, where the user device is an embedded vehicle device operating multiple IoT sensors, video streams and information and infotainment services.

The user now moves at 50 km/h to 110 km/h between coverage cells every 30 seconds to 2 minutes depending on deployment of Macro and Small cell Inter-site Distance (ISD) and actual vehicle speed.

This use case includes the following service and network referenced use cases from Annex A: (all from case 1 plus 33, 51, 53, 55 and 56).

8.3.4.6 Case 6: Fast, Multi-Access, Session & Bearer connection, with Macro HO

As for case 2, where the user device is a passenger operated personal mobile device in a vehicle moving at between 50 km/h to 110 km/h between coverage cells every 30 seconds to 2 minutes depending on deployment of Macro and Small cell Inter-site Distance (ISD) and actual vehicle speed.

This use case includes the following service and network referenced use cases from Annex A: (all from case 1 plus 53, 55 and 56).

8.3.4.7 Case 7: Fast, Multi-Access, Session & Bearer connection, with Macro HO

As for case 6, where the user device is a passenger operated personal mobile device in a high speed vehicle e.g. High speed train (HST) moving at typical speeds of 300 km/hr between coverage cells every 5 to 30 seconds depending on deployment of Macro and Small cell Inter-site Distance (ISD) and actual vehicle speed.

This use case includes the following service and network referenced use cases from Annex A: (all from case 1 plus 29, 53, 55 and 56).

8.3.5 Scenario Targets

Table 5 details the KPIs for improvement of this Scenario as a result of the development of NGP's.

Table 5: KPI's for Scenario - 03

KPI Name	Description	Units	Current Min Value	Current Max Value	Target Min Value	Target Max Value
SI(CP)	Service interruption, Control Plane gap	ms	E.g.: LTE 100 ms	E.g.: LTE 3 s	50 ms	500 ms
SI(UP)	Service interruption, User plane gap	ms	E.g.: LTE 100 ms	E.g.: LTE 3 s	50 ms	500 ms
HO_Rf	Handover Failure Rate, when the whole compound connection is lost. As a percentage of N handovers performed in unit time for a given scope (Cell, LAC, TAC, Network)	%	0,5	3	0,25	1

8.4 Multi-Access Support (including FMC)

8.4.1 Model Architecture

This multi-access and FMC section focuses on identifying key scenarios for leveraging existing 4G (LTE) networks and in particular the 5G network characteristics defined in 3GPP TR 23.863 [i.17] SMARTER eMBB (high data rates, low latency, high density, wide area coverage & low mobility) and 3GPP TR 22.864 [i.18] SMARTER NEO (network slicing, efficient data plane & content delivery, broadcast/multicast, policy control & charging, high availability and security) to enable combined use of fixed broadband (e.g. FTTx/xDSL) access and New Radio (NR) access networks.

8.4.2 Scenarios

Today a large number of residential and business users rely on fixed broadband (FTTx/xDSL) technology and MBB technology for accessing public Internet and private networks for their day-to-day use. Users should be able to intelligently combine both fixed and mobile access in a number of ways to meet their future needs.

In the scenario where users need ultra-high data rates for future applications, traffic originated to/from their devices in the home or in the office should be able to send/receive across both fixed broadband and mobile access (LTE/5G) either simultaneously or individually. It may be possible for the 3GPP system to specify network control policies to manage fixed broadband and mobile access as primary/secondary access type and dependent on the type of applications, time-of-day, location, type of user, type of end device and state of the network. User traffic may be operator provided services, customer's own services (e.g. corporate) or third party services (e.g. OTT).

In the multi-access scenario it is envisaged that such Access Context information that may be provided by the network and the user or user equipment will be used by the network to best serve them (human or machine) according to the context that the network can readily and securely access (see clause 8.5 on Context Awareness for further details).

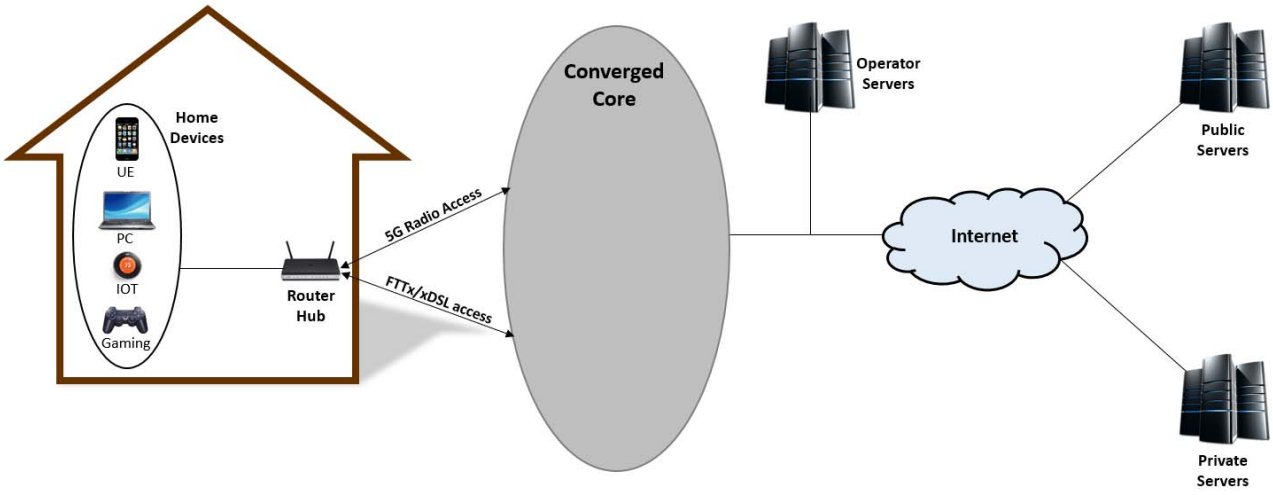


Figure 30: Multi-access and FMC scenario

In additional to a single operator scenario, the solution should also include the ability for two separate operators - Operator A providing fixed access and Operator B providing mobile access (with a relationship between them) to provide a joint service offering. This will place particular requirements on Next Generation Protocols.

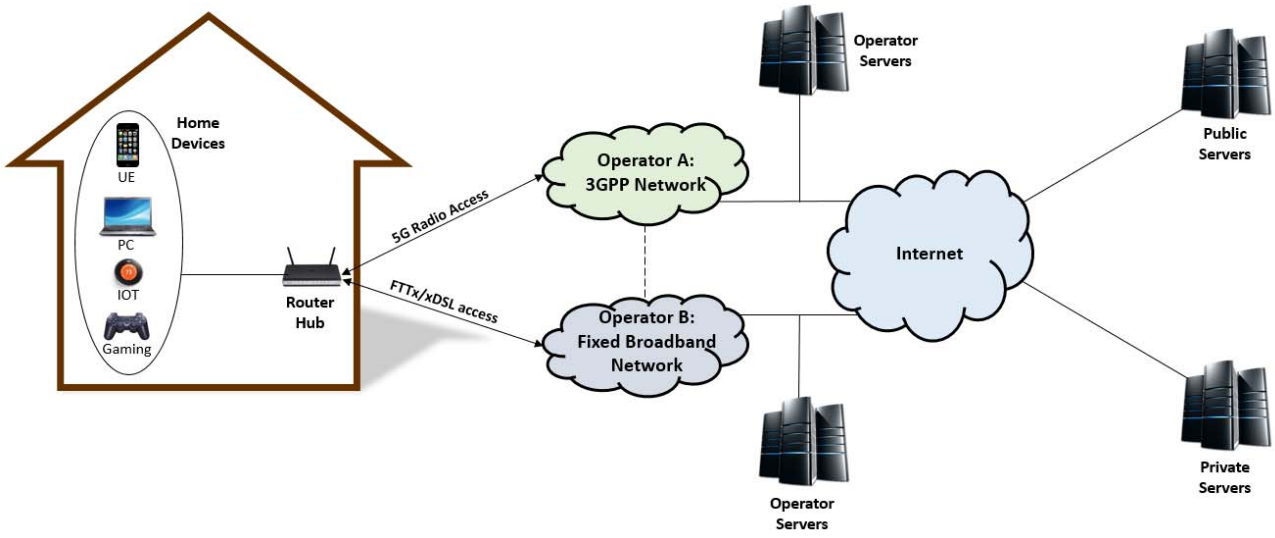


Figure 31: Multi-operator, multi-access and FMC scenario

8.4.3 Scenario Description

This clause further describes the support required for the models in this scenario:

- **Traffic Scenario 1 - Simultaneous use of mobile and fixed broadband access:** This traffic scenario allows constant and simultaneous use of both fixed broadband access and NR access in such a way that the end result data rate is close to the sum of both fixed and 5G data rates (> 90 %). It is assumed that a common implementation of this functionality would be to only use the cellular access when it is required i.e. for traffic peaks, as a load-balancing mechanism, or as a top-up as per operator policy.

- **Traffic Scenario 2 - 5G access as bandwidth boost:** This traffic scenario allows the on-demand use of mobile access to provide bandwidth boost to fixed broadband access. Users can trigger the bandwidth demand according to type of applications, time of day and type of Ues. This allows users to control their tariffs and increase their QoE. For Operators it enables the ability to increase revenues by upselling dynamic and temporary on-demand bandwidth for a specific time duration. As a minimum, users should have the ability select on-device turbo boost purchases via a smartphone app/webpage portal.

- **Traffic Scenario 3 - 5G access as failover:** Use of mobile access as failover mechanism in the case where fixed broadband goes out of service.

- **Traffic Scenario 4 - 5G access as fast provision:** Use of mobile access as fast provision service whilst users wait for their fixed broadband to be deployed or activated.

- **Traffic Scenario 5 - Symmetric Bandwidth:** This traffic scenario provides the end user with the same high data rates in the upstream direction as in the downstream direction. The faster uplink speeds can be used for cloud services, media and photo upload, etc.

- **Traffic Scenario 6 - Multi-Radio Access:** This traffic scenario provides the end user (human or machine) a combined high speed data service towards the network by combining as many available radio access technology options as possible for the users context in order to render the highest composite bandwidth service.

Alternatively, this scenario may operate multiple radio access technologies at the same time with each technology providing differentiated services according to their capabilities.

It is envisaged that in NGNs, there will be equipment that will be able to support several radio technologies at once combining N of the following different RATs at the same time and on the same user equipment, as follows: IEEE™ Wi-Fi™, 3GPP Cellular RF, 3GPP Cellular mm-Wave, Bluetooth™ and the dominant IoT radio access technologies at the time.

8.4.4 Applicable Issues

Issue-01: **Omission of FMC at Standards:** Currently there are no standardized architectures and/or methods for concurrent multi-access and FMC management at the access or core levels.

Recommendation-01: NGP should be able to deliver the aggregate throughput and speed of the FTTx/xDSL and multiple radio access technologies in both UL and DL directions to all traffic types and 3rd party applications.

Recommendation-02: The NGP should support mechanisms to ensure that combined traffic flows are delivered in sequence (to the end application) despite use of multiple radio access technologies and fixed broadband access being operated in combination.

Recommendation-03: The NGP should support a suitable addressing scheme to enable the combined use of multiple radio access technologies and fixed broadband access.

Recommendation-04: The NGP addressing scheme should enable multi-operator provision such that multiple radio access technologies and fixed broadband access could be provided across different networks.

Recommendation-05: The NGP should support the use of common user equipment (e.g. router hub at customer premises) that support multiple radio access technologies and fixed broadband access.

Recommendation-06: The NGP shall support dynamic and static address allocation to the common user equipment over multiple radio access technologies and fixed broadband access.

Recommendation-07: The NGP should support all traffic types over the combined use of multiple radio access technologies and fixed broadband access.

Recommendation-08: The NGP should support deployment multi-access scenarios with high with high data rates (10 s of Gbps).

Recommendation-09: The NGP should support multi-access deployment scenarios with ultra-low latency to enable real-time applications.

Recommendation-10:The NGP should support the option for Operators to provide the same level of security over the FTTx/xDSL link as is provided over the multiple radio access technology links.

Recommendation-11: For customer premises multi-access support, the NGP should provide a generic protocol between the multi-access aggregation point in the customer premises (e.g. router hub) and the aggregation point in the multi-access converged core that has a minimum overhead that can handle per packet or per flow scenarios.

Recommendation-12: The NGP should support a flexible (programmable) geographical distribution of the functional elements in the converged multi-access core and the operator services platforms, allowing for the FMC/Multi-access scenarios described in the present document to be supported on any geographical deployment.

8.5 Context Awareness

8.5.1 Model Architecture/Protocol Stacks

This clause details several context awareness scenarios where network functions are context awareness enabled with meta-data from other network functions and/or user equipment in order to drive network optimization and/or network organization and/or user QoE optimization.

The network model illustrates both fixed and mobile user equipment types, the considered virtual network function (VNF) itself and other virtual/physical network functions in the same network providing meta-data context information to the considered VNF as flows 1a, 1b, 1c.

The considered VNF then collates this information into a meta-data store for future/historical use at the same time as feeding the information direct to an associated SON algorithm, e.g. Automatic Network Organization (ANO) in order that it can control either itself and/or other network elements for the purposes of network SON for all connected users to this served slice of the network and/or UE1, UE2 QoE optimization on a per user basis as the control flows numbered 2. This concept is illustrated in Figure 32.

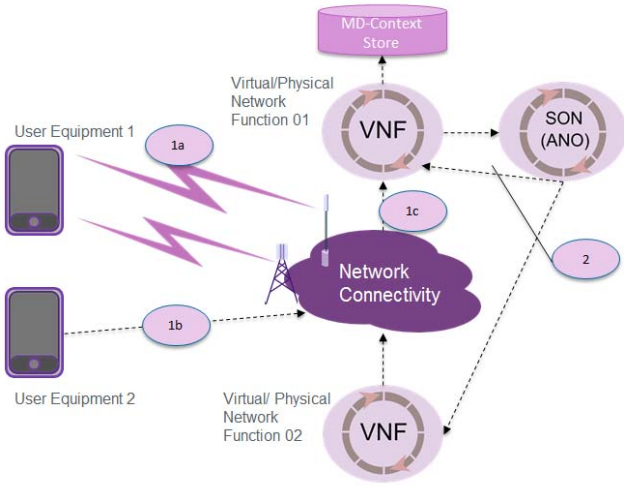


Figure 32: Context Enablement Architecture

8.5.2 Scenario Description

Most current communications systems traditionally operate with protocols that support either network control or user data transport over further protocols that provide network services. OSI layer 4 (L4) Protocols that support network control are commonly referred to in the ITU-T and 3GPP as Control Plane (CP) and protocols that support data transport are similarly referred to as User Plane (UP) protocols. These L4 protocols are supported by network protocols that are also further referred to as Transmission Protocol stacks in the same SDOs.

As communications systems evolve, specific information sets or information object class IoC definitions have been introduced, usually transported in the "control plane" to provide additional functionality to basic access and data transport in a piece-wise fashion to support network functions such as those following:

1) **Content Optimization** re-arrangement across a network.

2) **Mobility Optimization (**appropriate selection of anchors(s)/access point(s)) per communication transaction.

3) **Network Orchestration** control across a network.

4) **User QoE Optimization** to match current user experienced access network (e.g. radio access and channel condition) performance and current user experienced transmission/network performance according to user context and requested user action using device intelligence/performance gathering.

5) **Traditional Network Configuration Optimization** tuning as management of configuration, performance monitoring, alarms/events, faults and software.

The performance of all of these functions is largely dependent upon the quality and timeliness of the input information which is all contextual for the network.

Also, even for basic data transport and access control whilst the base communications system operates fairly well with limited essential information over the CP and UP L4 protocols, there is much research in recent years that has proven that additional contextual information used to support optimal configuration of these protocols and their associated functions, improves the Quality of Experience (QoE) offered by the network to all users generally and can be used to improve QoE to individual users.

As such, there exist many scenarios for communications networks where contextual information can be used to improve organization of the network and optimize QoE performance for its users.

However, currently, the information needed to drive these algorithms is difficult to obtain in a timely manner e.g. see Network Operation and Virtualisation Scenarios of the present document.

Therefore, in Next Generation Protocols there is a need to support a standardized and extensible framework to **efficiently provide** in a **timely manner** the context information that key network optimization functions need to optimize commonly experienced scenarios.

Some examples of scenarios that benefit from contextual information are outlined following:

Example 1: User Context for Content Optimization

Today there are many ICNs and CDNs that use proprietary algorithms to determine **what** contents and **where** in their networks they should cache/pre-load and **when** for best overall service across their customer base. When these networks are deployed across different access networks the performance differs significantly depending on **what** access type. The ICN/CDN operators already collect information on the history of the content items that people download and the popularity of each content item. However, such CDN/ICN networks are often blind to other information such as **where** the user is and **what** access network they were camped on at the time they downloaded the content.

To make the next step to optimize their network, these ICN/CDN operators need the Users to be able to provide additional context information about such parameters as Access Technology (throughput and latency capabilities), User Location (latitude/longitude, proximity to closest content server, locale = work, home stadium) to decide where to put content for this user/group of self-similar user types for improved QoE and what format to best serve according to device capabilities and quality of the content according to what this user/group of users will pay for the content at this location.

Example 2: User Context for Mobility Optimization

Mobility is the key differentiator of a cellular network as compared to fixed networks. Mobility is currently defined for all radio access technologies in terms of "best candidate cell". However, the real mobility scenario is more complex, where a user has access to multiple access technologies at once and sometimes multiple anchor/access points within each access technology (e.g. Cells, SGW, AP, etc.) within supportable access performance bounds (e.g. multiple candidate cells that the user could connect to at this time). However, currently, the network does not react to user provided context information on **what** they are trying to do or **where t**hey are or **which** device type they are (other than the access capabilities of the device).

For example, on a Radio Access Network (RAN) frequently the access technology drives the user to the "best cell" when they do not really need the best candidate cell for the transaction they are requesting or they simply connect to the a sub-optimal heterogeneous cell level for the transaction they are requesting as they ignore the user context.

This kind of scenario limits the best use of available access technologies and their channels. Clearly, better use of each access network could be made if the network was context enabled and knew the context of the user on a "user by user" and "request by request" basis. Selection of appropriate Anchor Point/Access Point would improve overall network efficiency for all and matching this with available network resources based on context would further improve overall network efficiency and load management.

When a mobile user attaches to a network today the cell or access point that they connect to is determined purely by the radio conditions that they can achieve at the time.

However, there are contextual reasons why this is not always the best solution for best served QoE.

To make the next step to optimize their network they need the Users to be able to provide additional context information about such parameters as Access Technology and Mobility/Entropy to decide where to put content for this user or type of user for improved QoE.

Today, if the user is a mobile broadband user and has LTE connectivity then the LTE system will select the best candidate cell purely on the basis of RF performance. If the Macro outside their window looks like a better candidate then they will select this even though the indoor system may only be slightly worse for their particular indoor location in terms of RF performance but offer much better throughput opportunities.

Alternatively, if the user is moving fast down a motorway it would seem inefficient to connect to any small cells as it passes the edges of dense urban areas as this would involve notable handovers as compared to remaining camped on a Macro at high speed.

Similarly, if there are many static IoT devices in a system over a wide area that need to collectively be gathered together as fast as possible then it would seem efficient to collate these measurement reports over that wide area at a Macro cell rather than each of the devices collect o any local small cell and then have to collate from there.

It would be better if a slightly wider range of candidate cells could be obtained to connect to as the user moves through the cellular deployment and then temper it with the context of the user in terms of what their mobility entropy is. To do this today a number of indirect measurements at the Access Network and complex handover algorithms are employed. However simple periodic supplementary context information on mobile device type, speed and higher level service requirements information from the user would significantly reduce the complexity of the access camping problem for best QoE.

Example 3: Network Orchestration & Self Organizing Networks (SON) Meta-Data

SON algorithms have been widely deployed in 3GPP UMTS and LTE/LTE-A communications systems, however the most successfully deployed algorithms are those that operate at the edge of the network with the tightest time constraints and with RF parameter input data as Distributed (D-SON) algorithms from the base stations operating at the access radio level and using input information that changes fast as compared to the rate of the information on the user data transport path. Successfully deployed algorithms now in the market at the D-SON level include Automatic Neighbour Relations and Automatic PCI allocation, supporting optimization of mobility in terms of neighbours and cell identity planning.

Algorithms that operate higher up in the cellular topology hierarchy, typically at the level of an OMC as extension to the OMC/Trace Servers/or in the Core Network have been less successful in the cellular market. The lack of success at the C-SON level is largely due to the lack of easily available and timely information to drive this level of algorithms, from OMC-PM, trace and other network based sources. At the C-SON level the algorithm is typically trying to respond to changes at the rate of diurnal/hourly traffic group changes to optimize the network.

Hybrid or H-SON is also noted in the 3GPP SON standards, as a set of algorithms that operate using data from both the D-SON base station/RAN level and the C-SON, OAM level together to meet timeframes for algorithms that can provide solutions for one or both of the organization/optimization timeline targets of C-SON/D-SON.

D-SON algorithms are often proprietarily implemented and have varying impacts on other aspects of base station performance and the way that mobiles operate to provide information feeds to the D-SON algorithm. For C-SON and H-SON algorithms are also proprietarily implemented and struggle to obtain current timely information from OMC, trace, UE agents.

In 5G with more simultaneously active Access Technologies in typical user environments, the need for SON is greater than ever, so it is becoming more important to be able for these algorithms to get timely access to input information to drive them. To do this, SON algorithms, wherever they are located in the network hierarchy, need access to a standardized structure of timely Meta-Data from both the networks themselves and the users.

In all of the user based QoE optimization examples currently being discussed as use cases for next generation access technologies, often the user is willing to share information with their provider to gain better QoE, but they do not want this to be done without losing their privacy control (non-proliferation of preferences, right for their information to be forgotten, etc.) and they also do not want the information sharing to notably impact their throughput.

Example 4: User QoE Optimization

At present in LTE cellular systems, the selection of which access technologies to employ to support any new service request by a user is largely left up to them and usually means choosing between one of 2 options:

i) 3GPP LTE; or

ii) Wi-Fi™.

Also, some mobile vendors support simultaneous use of both i) and ii) access technologies using MP-TPC. However, for LTE, this type of solution has limitations in the market of today on implementation and practical deployment realizations. These typical limitations include aspects like:

i) being mainly being optimized for downlink transmission; and

ii) being inhibited by the firewall settings used by some operators.

Most mobile users are now fluent in managing these two access technologies on their devices, but tend to operate access selection settings purely based on perceived cost for all transaction types per technology rather than a potentially more efficient per transaction type which is also time consuming to manually change access settings per transaction type.

For next generation access systems, users will have many more access choices on their devices than included today and some access types will be much more suitable for specific communication transactions than others. It is expected that in next generation networks, at least two and potentially all of the following access technology options will be available on mobile devices:

- 5G-Cellular;
- 5G-mm-Wave;
- Wi-Fi™; and
- Fixed infrastructure.

So, in the next generation era, the user will need a more automated approach to access configuration and it will need to be more dynamic to be able to optimize the users overall QoE across these multiple access technologies on a per communications transaction or per group of self-similar communication transactions.

To best optimize QoE in next generation, and select the best access options per communication transaction, a standard format of meta-data and meta-data exchange is required so that either user device based or network based QoE algorithms can assist the user to provide best QoE across the currently available and usable access technologies at each transaction initiation.

The kinds of meta-data expected to be required to drive access selection management for best user QoE are as follows:

i) communications transaction request description;

ii) available access technologies at the device;

iii) current user assessment of the likely performance of each access technology based on user location and RF performance relative to the access point or connection.

It is expected that the actual composite access decision, ideally would use meta-data from both RAN and UE entities.

Example 5: Traditional Network Configuration Optimization

The NFV and SDN fields are rapidly evolving to virtualise network architecture and management of topology optimization with the new field of orchestration in the form of MANO, as specified by standards bodies such as ETSI ISG NFV and the ONF. This evolved, virtualised and orchestrated approach to networking requires XML based scripted control to all entities with suitable interfaces.

To integrate optimization and organizational orchestration algorithms into next generation networks then NGP needs to drive them with efficient and timely Meta-Data. For configuration this can be on a fairly slow basis, but with the advent of network slicing and the ever present need to support critical communications robustly then network slices need to react to network events and performance meta-data much faster than in traditional OAM systems driven purely with FM/PM/CM statistics.

As such there is a need for meta-data to drive NFV/SDN/MANO using user and network generated information in a timely and efficient manner.

8.5.3 Applicable Issues

Issue-01: **Meta-Data Context Transport:** Today, there is no standard network protocol available today that is optimized for meta-data transfer to support network functionality/capabilities such as, network organization, network optimization and user QoE optimization.

Recommendation: In generating a suite of protocols for next generation networks there should be an accommodation to provide for the support of transfer of a standard extensible data structure of meta-data Information Object Classes (IoC) that can be readily interpreted by devices and functional network entities specifically for the purposes of optimizing and organizing the network and optimizing user QoE.

Issue-02: **Meta-Data Context Structure:** There are a plethora of modern protocol data structures that could be used to support either bit-oriented, byte oriented or structured English syntax oriented, for example XML. However, most of these algorithms need context data at varying levels of timeliness and efficiency.

Recommendation: NGP should provide a meta-data protocol with structure that would support dynamic protocol formatting, with a simple fixed IoC, TV data structure that supports both fixed common IoCs for each version with extensible TLV coding in any of these formats, indicated with a TV field. For extension these TLV coded fields can be migrated into TV format depending on protocol evolution following common usage. It is anticipated that the meta-data protocol could be built using a standard template architected transmission protocol from the Next Generation Protocol activity and then adapted to support the proposed standardized Meta-Data fields and IoCs.

Issue-03: **Pre-Defined IoC Support:** In order to efficiently support the development of algorithms that support the kind of scenarios detailed in this clause as examples, there are certain standard IoC types that are required in order to enable meta data required to drive the organization and optimization of next generation networks.

Recommendation: The following typical set of pre-defined and common conditional contextual meta-data IoCs should be included in a next generation protocol supporting efficient and timely meta-data/context information transport:

User Originated

What:	Equipment Capabilities, Access Capabilities, Content history
Where:	Address[Locale, Current Location(Cell/AP/Connection point, Latitude/Longitude, TAC)], Entropy history, Mobility history, Speed, Heading
When:	Current Access opportunities, Recent Access performance assessments per access type
Why:	Recent Access Failures
Who:	Name of Equipment, Name of User Communication, Transaction History, Type of User

Network Originated

What:	Network Function Type
Where:	Address[Locale, Current Location(Cell, Latitude/Longitude, TAC)]
When:	Current Performance, load statistics, Collective Access Network performance/History per type of user Current Alarms
Why:	Recent Access Failures/History per type of user
Who:	Name of Function

Issue-04: **Meta-Data Scope:** In order to keep the transport of contextual Meta-Data efficient then the consumer of the Contextual Meta-Data needs to be able to discover potential suppliers and be able to select which IoCs they can deliver.

Recommendation: In defining a new Meta-Data Protocol, there should be a method for a consumer to discover a supplier setup a stakeholder relationship with them and be able to agree which IoCs they are going to provide on an ongoing or periodically updated basis or on a one time basis.

Issue-04: **Privacy and Trust: Meta-Data Scope:** In setting up any relationship between two network entities whereby one provides to the other to consume, there needs to be some protection afforded to the stakeholders in the relationship to ensure mutual trust and privacy of the data.

Recommendation: In designing a suitable meta-data protocol for context transport between network entities, there should be a set of security procedures associated with the establishment of the relationship and the subsequent data exchange including facets such as mutual authentication, data integrity and optional encryption.

8.5.4 Applicable Use Cases (from Annex A)

SMARTER referenced use cases: 2, 5, 6, 7, 8, 9, 26, 27, 28, 30, 34, 35, 36, 37, 38, 47 and 51 from SMARTER ref in Annex A are identified as having notable Context Enablement scenario dependence and/or are required in order to support next generation Context Enablement enhancements.

8.5.5 Scenario Targets

The following target KPIs in Table 6 are described for use by NFs to optimize or organize within a given type of timeframe.

Table 6: KPI's for Scenario - 05

KPI Name	Description	Units	Current Min Value	Current Max Value	Target Min Value	Target Max Value
Trel(10, UE-NF)	The time to setup a metadata context relationship between a user equipment supplier and consumer entity for an IoC list of 10 IoCs when operating at the per user QoE optimization timeframe.	Ms	N/A	N/A	200	500
Trel(10, NF-NF)	The time to setup a metadata context relationship between NF(01) supplier and NF(02) consumer entity for an IoC list of 10 IoCs when operating at the per user QoE optimization timeframe.	Ms	N/A	N/A	500	1 000
Ttrx(10, UE-NF)	The time to transfer 10 metadata context IoCs between UE and NF when operating at the per user QoE optimization timeframe.	Ms	N/A	N/A	10	100
Ttrx(10, NF-NF)	The time to transfer 10 metadata context IoCs between NF(01) and NF(02) when operating at the per user QoE optimization timeframe.	Ms	N/A	N/A	10	100

Where:

i) L2 transmission rate over the access technology for the MDP is assumed to be better than 10 Mbit/s.

ii) L2 transmission rate over the network technology for the MDP is assumed to be better than 100 Mbit/s.

iii) It is assumed that the mean length of an IoC is 100 octets when coded in octet format.

8.6 Performance Improvement & Content Enablement

8.6.1 Model Architecture

This scenario identifies the current issues that limit transmission performance when there is an access network in the end-to-end (E2E) path. The scenario also identifies the issues and challenges with the enablement of smart content management at the mobile network edge. These limitations are especially apparent with respect to the limitations that are introduced by TCP/IP suite. Furthermore, the potential features of evolved network functions at the network edge that could enable smart content management are envisaged.

A high-level illustration of the typical multi-access user equipment (UE) transmission scenario is presented in Figure 33. The UE may be connected to the core network via wireless access network (e.g. LTE, Wi-Fi™, etc.) or fixed broadband access network.

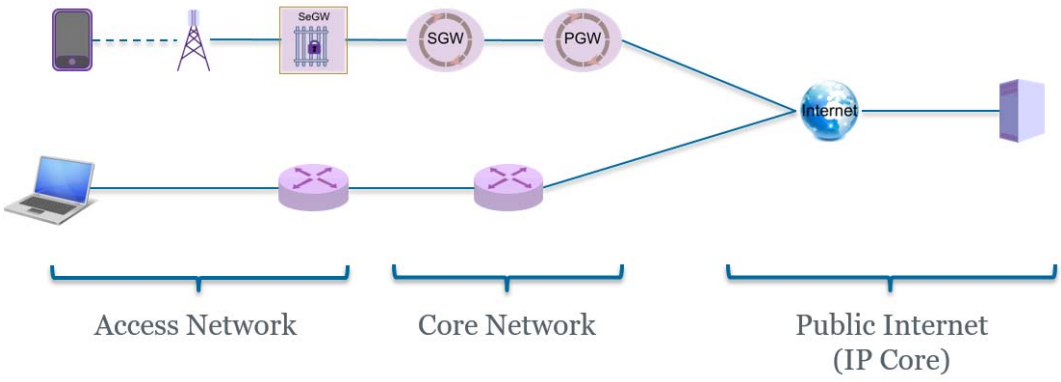


Figure 33: Conventional content delivery architecture based on TCP/IP suite

It should also be borne in mind that although this simplified model of internet access for fixed and mobile is sufficient to address performance issues by Access, Core and Internet segments of an ETE path, often the real world path is much more complex involving many more separate entities that an ETE performance path has to traverse to provide internet access to an end user, as illustrated in Figure 34.

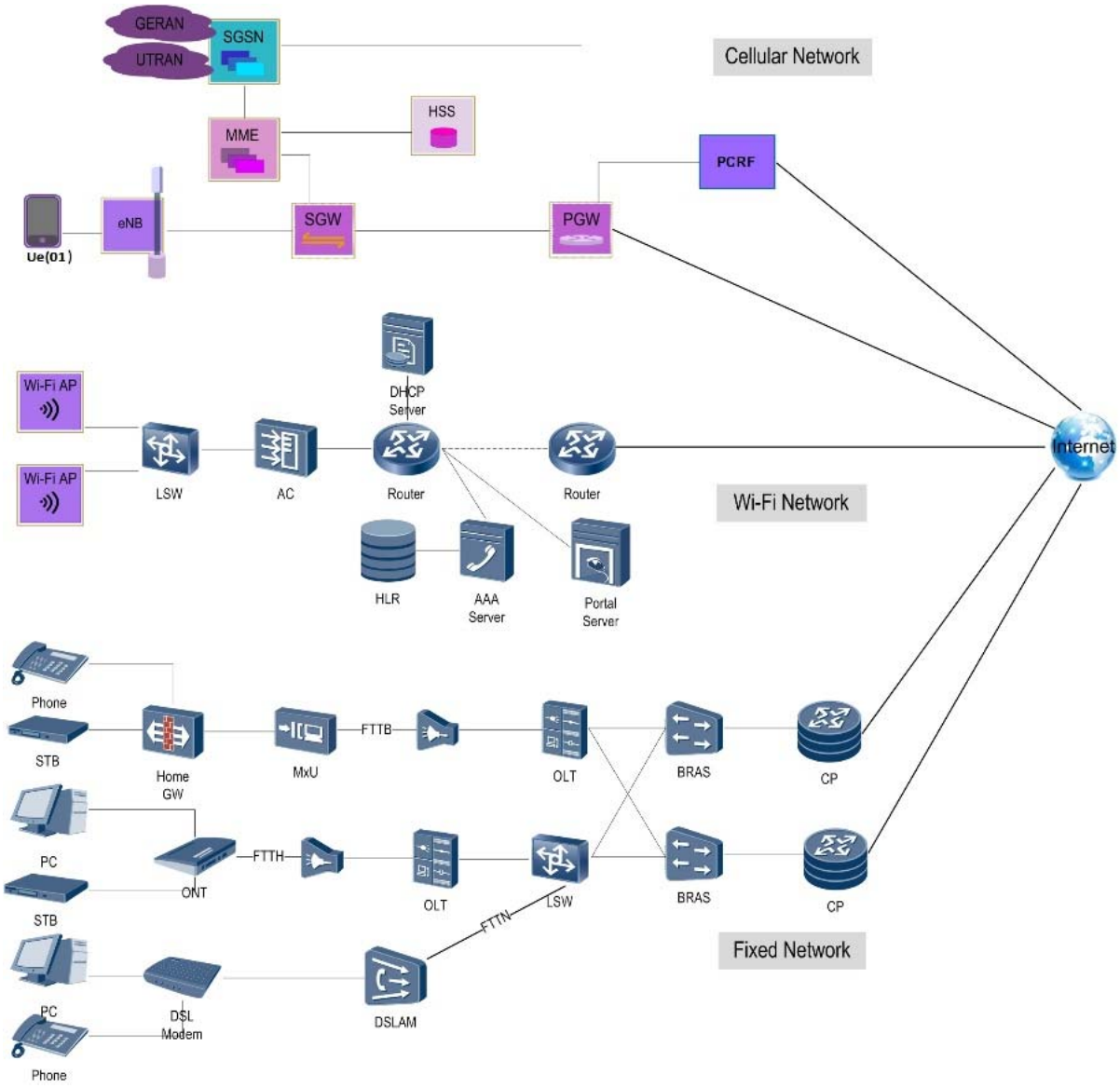


Figure 34: ETE Heterogeneous (Fixed/Wi-Fi™ and Cellular) Internet Access Path Complexity Examples

A detailed illustration of the protocol stack involved in this scenario is presented in Figure 35. Note that Figure 35 shows only the scenario with radio access network only - a fixed access network has a simpler protocol stack and is not illustrated in the present document.

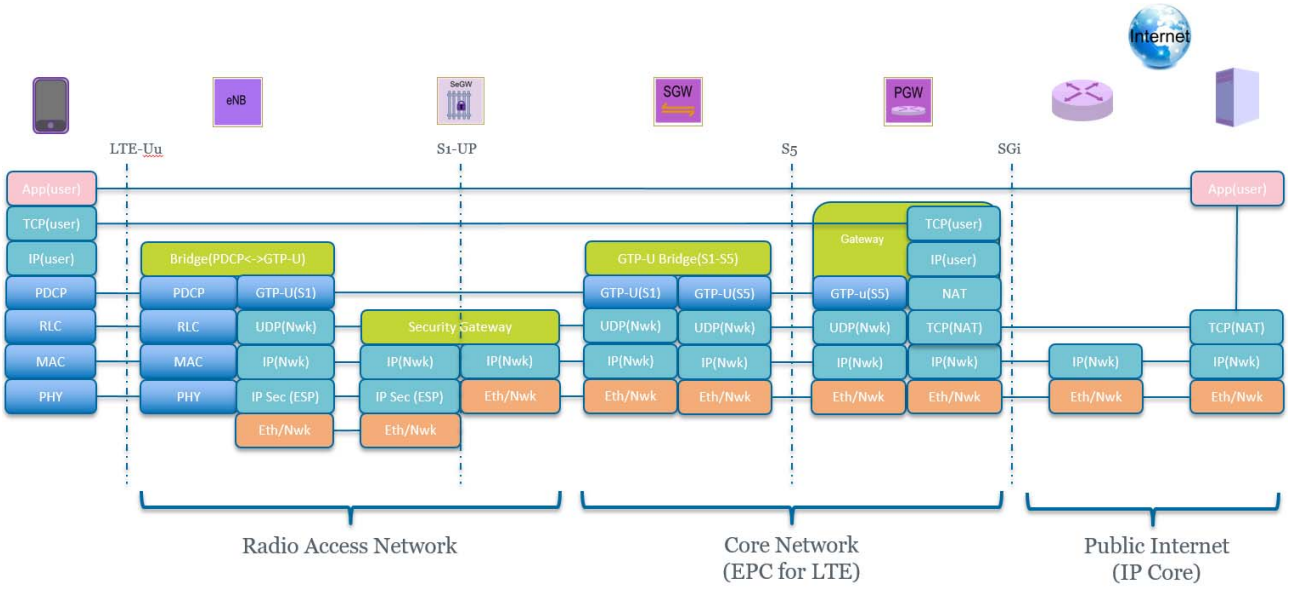


Figure 35: Protocol Stack Illustration of a Typical Content Delivery Path with RAN

In this scenario, when a UE requests a piece of content, the request and the content may be delivered via multiple access connections, such as Wi-Fi™, LTE, fixed network, etc. Regardless of the type of access network that is used, an E2E connection will be established between the UE and the server under the TCP/IP protocol suite. The UE's request is normally resolved to a fixed server IP address as instructed by a DNS server.

It is an important observation that TCP has become the de-facto protocol for supporting content applications, e.g. web-based applications and video streaming applications. In this context, when an E2E TCP connection is established between UE and server, the following factors often degrade the download throughput:

- Handshake: the handshake process means extra delay during setup of a TCP connection, which especially affects short and bursty network applications such as webpage loading.
- Slow start: TCP uses slow start as part of its congestion control mechanism to probe network condition, which means download throughput is especially low for small file downloads, such as in webpage loading and video streaming sessions with small segments/chunks. More details on the performance issues of congestion control are discussed in clause 8.6.3.
- If network latency fluctuation and/or packet loss take place at the core network and/or public Internet, the entire UE-server connection suffers in terms of throughput due to TCP retransmission/out-of-order packets, even though the access network side may be perfectly fine. This is caused by TCP congestion control mechanism.
- Mismatch among buffer management of different layers: there are 3 separate buffer management schemes operating at RRC, PDCP and TCP layers on their own, and they may conflict with each other. Such conflict often causes latency fluctuation in the radio access network.

8.6.2 Scenario Descriptions

8.6.2.0 Introduction

This clause describes the support required for the models in this scenario.

All of the Performance aspects need to be supported in context, which includes both "static" operation and Mobile operation where the requirements of the Mobility Scenario have to be met concurrently.

Each performance scenario should be treated whilst bearing in mind that a user may be operating multiple applications on an End-to-End basis at a time with sufficient throughput and latency to support each application according to the various kinds of access network used to provide the edge link(s) of the path.

For all of the scenarios included here, currently internet access paths that include radio access links all require improvement for next generation networks.

For future applications it is expected that it will become common-place in next generation networks to see UHD video streaming application (e.g. 4K/8K/VR). When using these video streaming formats then the performance required from the supporting transport protocols needs to significantly improve as compared to today's TCP/IP protocols.

The scenarios 1 to 4 are used to discuss TCP performance limitations in real Internet applications.

8.6.2.1 Scenario #1 - Adaptive video streaming

In widely-adopted adaptive streaming protocols (such as MPEG-DASH and HLS), a video is divided into multiple fixed-length segments at its source, where the typical segment lengths are 2 to 10 seconds. Take the example of DASH - when a user starts watching a video, it first downloads an XML file called MPD (Media Presentation Description) of the video, which contains its metadata and URLs to each of its segment. The UE then requests the video segments for playback. During a playback session, the UE's request pattern is different depending on how much video buffer is available on the device:

- If the buffer is low (e.g. at the beginning of a playback session), the UE will start requesting the next segment as soon as the previously-requested segment is delivered. This will last until the UE's buffer is saturated (e.g. 30 seconds).
- If the buffer is saturated, the UE will request the next segment only when the buffer falls below a certain threshold (e.g. 30 seconds), so that the buffer can be maintained above a certain level. In other words, there will be a predictable gap between each of the UE's requests, where the gap equals the length of each segment.

At the beginning of a video streaming session, the UE would query the DNS server for the IP address of a content server. Based on the resolved IP address of the content server, the UE then downloads the MPD file from the server and all the remaining segments as instructed by the URLs in the MPD (until the playback stops). Note that since the MPD file is only downloaded once, as soon as the playback starts, the UE cannot choose which server to download the video segments from.

In practice, the network condition fluctuates significantly in both access and core networks. Specifically, in a radio access network, the network latency often fluctuates significantly due to the mismatch among buffer management schemes at different layers, while error control schemes like HARQ and FEC maintain packet error at a relatively low level. On the other hand, in the core network and public Internet, the network latency is much steadier than in wireless networks. However, the packet error rate is higher due to the buffer overflowing at routers along a content delivery path.

It is well known that the predominant network latency fluctuation is generated in the access part of the ETE path (for example in a radio access network), whilst packet errors (e.g. lost, out-of-order or duplicate packets) may be introduced by either the radio and/or fixed network parts of the ETE path. However, since the E2E TCP connection is established between the UE and the content server (as resolved by DNS server), and the UE-server path includes both radio and fixed networks, the server end's TCP congestion control mechanism will take into account both latency fluctuation and packet error in the path, which causes smaller congestion window that affects throughput significantly.

Furthermore, at the beginning of each video segment download, TCP slow-start mechanism will be employed. For smaller video segments (e.g. < 5 MB), this means the download will be finished before the congestion window has become large enough to utilize the entire available bandwidth. As a result, the allocated bandwidth is wasted and the download throughput is lower than desired. Note that this issue applies to both request patterns mentioned above - even when only one TCP connection is opened and maintained throughout a DASH streaming session by default. This is because there is a gap between each segment download, the server regards the TCP connection to have become idle and ignores the congestion window from the last download session.

8.6.2.2 Scenario #2 - 8K Video Streaming

8K UHD refers to the horizontal resolution in the order of 8 000 pixels across the image, forming a total image with dimensions of (7 680 × 4 320 pixels), with typically each pixel occupying 12 bits for colour memory space. The frame rate of a basic 8K video stream is 60 fps, which means there are sixty frames within 1 second of video content. Thus, the uncompressed bit rate of an 8K video stream can be computed as: 7 680 × 4 320 × 12 bit × 60 fps ~ 23,89 Gbps. H.265 compression is a currently widely employed and high efficiency video codec. It normally provides a compression ratio of about 200 times. So, the required throughput of a compressed 8K video can be approximated as 23,89 Gbps /200 ≈ 119 Mbps.

In the transmission of a video stream, there are always fluctuations in the process of transmission which can severely degrade performance. So, typically an additional 50 % extra bandwidth is required for a channel supporting such a stream on top of the raw stream bit rate, to maintain ETE video quality. In this case, the actual requirements can be computed as: 119 Mbps × 1,5 ~ 178 Mbps to support 8K UHD transmission. 'High Quality' 8K video transmission usually demands an even higher frame rate (e.g. 120 fps) to provide a better user experience, in this case the user demand throughput is doubled, as follows: 178 Mbps × 2 = 356 Mbps.

8.6.2.3 Scenario #3 - Live Virtual Reality

To ensure a good user experience, Live Virtual Reality, a much higher resolution than 8K video is required. For instance, ultra VR required 16K × 16K resolution as well as a 90 fps refresh rate. For a 2D scenario with H.265 coding, the video data rate for a 2D Range of Interest(ROI) may reach up to 16K × 16K × 12 bit × 90 fps/200 ~ 1,38 Gbps and thus the required bandwidth, allowing for typical communications path throughput fluctuations can reach up to 1,38 × 1,5 ~ 2 Gbps. For a 3D scenario, the bandwidth requirement may increase much higher due 3D full view requirement and an even higher refresh rate.

In addition to the high bandwidth requirements of VR, there needs to be a supporting transmission network to provide a communications path with low latency as well. Specifically, there exists a Motion to Photon (MTP) latency requirement for VR applications of less than 20 ms. The 20 ms MTP consists of several parts, including motion capture, coding at the server side, video streaming and reorganization at the server side, network transmission delay, decoding time on user-end, screen response time and refresh time. Each of these component function parts will take a certain time. So, the time left for network transmission is typically less than a quarter of this value ~ 5 ms. This stringent VR latency requirement is a challenge to existing networks where for example in LTE-A today just the air interface link of an ETE internet path is typically much more than 5 ms.

8.6.2.4 Scenario #4 - URLLC For Time-Critical IoT

Ultra-Reliable, Low Latency (URLLC), Internet-of-Things (IoT) is another important use case scenario-set that highlights requirements that are difficult to satisfy with existing internet connectivity paths where a part of that path includes a radio access link. A common example of this type of scenario is that of time-critical sensor-actuator systems such as are currently supported by bespoke SCADA systems. Here a control loop would scan up to 300 remote stations with sensors and expect to control any of the actuators that need adjustment in that scanned set within a few seconds. So that on average the control cycle per remote station could be anywhere from less than 1 ms to 10 seconds depending on the criticality of the determined actuation scheme and which remote station to control first.

These kind of close-loop control systems borne over heterogeneous communications networks have very low latency requirements for the E2E network connecting the sensor and actuator.

8.6.3 Issues with TCP Congestion Control

8.6.3.1 An appraisal of Congestion Management

A number of problems have been identified that some access technologies, e.g. Cellular, have in terms of the way in which they interact with the Internet. Many of these problems have to do with packet loss, and congestion. The problem is that TCP cannot distinguish loss due to congestion from loss due to the nature of the media. This problem is well known in the industry.

There are two approaches to congestion management:

i) congestion Control, constantly testing the cliff of congestion collapse; and

ii) congestion Avoidance, constantly testing the knee of the curve to optimize the trade-off between response time and throughput.

NOTE 1: See Figure 36.

NOTE 2: The term congestion 'cliff' is the point at which a session collapses.

NOTE 3: The term 'knee' is the point at which a session begins to notably deteriorate.

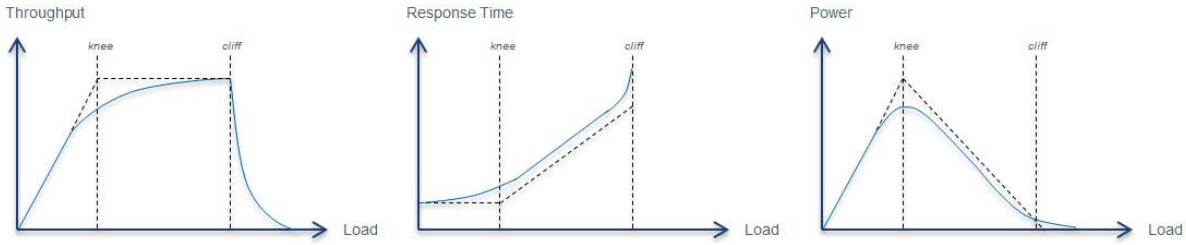


Figure 36: Congestion Curves

TCP congestion control detects congestion by constantly increasing the number of packets sent until loss is detected. Hence, TCP congestion control is always testing the edge of the congestion cliff and works by actually causing congestion and the loss of data.

Given the increasing diameter of the Internet (typical path length) that is now considered to be about 20 hops, and the increase in diameter contributed by the mobile networks (can be as much as 50 % - 100 % greater) this means that the delay in the congestion response further increases with more loss. With TCP testing the cliff, the longer the response to congestion increases the probability of going closer to the edge or even going marginally over the edge and having to recover makes the loss worse.

The problem of congestion in connectionless networks has been thoroughly investigated by Jain, see [i.11] and [i.12] and found that optimizing for the knee of the curve, rather than the cliff, provides the best approach to optimizing response time and throughput while minimizing packet loss. Jain's work shows that explicit notification (ECN) is essential, not only because congestion is not the only cause of lost packets and therefore not a good indicator of congestion, but because Jain proves the low-pass filter for congestion notification should begin when average queue length is *greater than or equal to one packet.* Far earlier than waiting for congestion to cause packet loss. Use of ECN also ensures that the effects of the congestion response are limited to the layer in which the congestion occurred and is not predatory as implicit notification is. Jain also proposes using the flow control window to slow the sender rather than a separate congestion window at the source. Not only is this simpler, but also allows the congestion response to be coordinated with flow control policy.

While there has been an effort to add ECN to TCP, it adds it to the existing mechanism without taking into account the other associated results. Furthermore, many TCP implementations fail if ECN-related bits are used. All of the current TCP schemes are variations on packet discard behaviour during perceived congestion, which actually creates the problem trying to be avoided.

The combination of ECN, notification when average queue length is greater than or equal to one, and optimizing for the congestion knee rather than the point of collapse indicates a network that would behave very differently than currently and is likely to solve most of the problems that have been raised. The saw-tooth behaviour of TCP congestion control thwarts effective QoS strategies in the Network Layer to reduce jitter. This in turn requires large amounts of buffering at the destinations to smooth it out. Again, following the results for [i.11] and [i.12] who finds that "one size does not fit all." Instead congestion strategies should be matched to QoS classes. There is therefore an opportunity to solve this problem as well by moving congestion management to where it was originally intended to be, in the network layers, not in the layer with the greatest scope, such as TCP.

8.6.3.2 An Introduction to Current TCP Congestion Mechanisms

TCP congestion control is the source of many performance problems in the Internet and may be one of the most severe, most fundamental problems confronting the Internet today. Congestion normally takes place when too many PDUs travel through the network and arrive at a peer at the same time, which leads to packet errors such as duplicate ACKs, out-of-order packets and hence, TCP retransmissions. It is worth noting that congestion is not necessarily due to the network being overloaded - it can take place in lightly loaded networks too. It also takes place in wireless networks due to the nature of radio air interface.

When congestion takes place in an E2E connection, the general solution is to reduce the transferred data volume in the hope of mitigating the packet errors. In early TCP implementations (e.g. Tahoe in 1988), the TCP transmitter follows a simple slow-start (with an initial congestion window - *initcwnd* of 1 TCP segment) and congestion control mechanism. The congestion window (*cwnd*) increases exponentially until it reaches slow-start threshold (*ssthresh*), which equals the receivers' advertised receiving window size. It then enters congestion avoidance phase, where the *cwnd* is increased linearly and slowly (instead of exponentially under congestion control).

TCP Tahoe's congestion control mechanism responds to packet errors as follows. For example, if three duplicate ACKs are received, or when a retransmission timeout (RTO) happens, TCP Tahoe sets *ssthresh* to half of the current *cwnd*, reduces *cwnd* to 1 TCP maximum segment size (MSS), and restarts the slow-start phase.

A significantly improved version, TCP Reno, was implemented in 1990. It introduces a "fast recovery" phase to improve TCP performance under retransmissions. When three duplicate ACKs are received, instead of setting *cwnd* to 1 MSS, TCP Reno halves the current *cwnd*. Meanwhile, *ssthresh* is also set to the same value, hence skipping the slow-start phase and directly entering congestion avoidance phase. Note that TCP Reno still set *cwnd* to 1 MSS and restart slow-start in the case of an RTO. Such "fast recovery" phase is further improved in TCP New Reno (IETF RFC 6582 [i.19]) in 2012, which introduces selective acknowledgement option (SACK) technique. Interested readers are referred to IETF RFC 2018 [i.20] for more details.

There are dozens of newer TCP implementations that are developed since TCP Reno, all of which employs different congestion control algorithm that adapts *cwnd* to the congestion events in different manners. For example, TCP Cubic, which is widely adopted in recent Linux kernels, uses a cubic function to adjust *cwnd* with respect to packet errors.

Without doubt, these improved TCP implementations (especially New Reno and Cubic) have significantly improved TCP performance in the Internet. However, it is important to notice that TCP is still a protocol that was originally designed for fixed networks, and most TCP implementations, including New Reno and Cubic, work best in networks with long latency and large bandwidth (i.e. large BDP networks, where BDP is bandwidth-delay product) where more bytes are in transit at a time. On the other hand, these schemes are not optimized for wireless network characteristics, especially in LTE networks where multiple hops and a complicated protocol stack are involved between the user and the public Internet.

It has been discussed earlier that in LTE networks, the main challenge for TCP at the access network side is variable latency caused by mismatched buffer management between layers and the nature of the radio air interface. Meanwhile, the average latency at the RAN is relatively short compared to the Internet (around 20 - 40 ms) except during the spikes caused by RAN buffer overflow (which can typically reach 500 - 600 ms). Also, each user's RAN bandwidth is of the order of 10 Mbps to 200 Mbps (depending on LTE category and the number of users in a cell), which is generally a lot lower than in the Internet (typically 1 Gbps+). Therefore, the BDP is considerably lower in the RAN than the wired internet, for which TCP was originally designed and optimized. Hence, in an E2E TCP connection which involves both RAN and public Internet, each with significantly different network characteristics, the existing TCP implementations cannot achieve satisfactory performance.

Clause 8.6.4 specifically discusses a number of applicable issues regarding TCP performance, and makes recommendations to tackle these issues.

8.6.4 Applicable Issues and Recommendations

Issue-01: **TCP Three-Way Handshaking:** The delay incurred by the three-way handshaking used in TCP to start every session adds latency to the session start-up, and this is unnecessary.

Recommendation: For all situations that require feedback mechanisms, NGP should adopt protocols that explicitly bound Maximum Packet Lifetime, the time to wait before Ack, and the time to exhaust retries in all protocols with feedback, see [i.13].

Issue-02: **Implicit Congestion Notification** can lead a protocol to miss-diagnose congestion and lead to unnecessary loss of data and unnecessary overhead to recover from the loss.

Recommendation-01: Improved forms of congestion, latency and PER feedback within known bounds that limit the variance of response time, time to notify and variance in time to notify, should be considered for NGP.

Recommendation-02: Explicit Congestion Notification (ECN) should be used for congestion notification with each layer that does congestion management. (It is likely that any layer that relays will require congestion management over the scope of the relaying).

Issue-03: **Transmission Configuration:** There is frequently no way for a TCP layer to be configured for a variety of different Access Technologies. TCP is supposed to manage over heterogeneous ETE paths but is actually inflexible in practice.

Recommendation: NGP protocols should consider transmission protocols that can be configured dynamically for a variety of access technologies according to policy.

Issue-04: **Level of TCP Congestion Control:** TCP Congestion Control operates above the layers that support it and impose QoS.

TCP actually thwarts the QoS control mechanisms at these lower layers and makes it makes it difficult to coordinate an overall congestion response to QoS.

This issue is particularly true when TCP is operated over a heterogeneous communications path that includes access links operating over channels with notably variable performance (such as the Cellular LTE standard).

Recommendation: Congestion management should occur in the same layer where QoS is enforced for that layer so that congestion management policy and QoS policy can be coordinated.

Issue-05: **Implicit Nature of TCP Congestion Notification:** Because TCP Congestion Notification is implicit, and has a longer response time than the lower layers that support it, it therefore has the greatest variance in response time, and it is predatory with respect to the lower layers. In other words, the slower response of the transport layer will work against faster response in the lower layers. The implicit notification means that a TCP response cannot be avoided unless there is no loss of data. Note that protocols that attempt to avoid loss by operating at the knee will be at a disadvantage with respect to TCP. TCP behaviour pushing the edge of the cliff, i.e. greedy, will starve flows attempting to operate at the knee. (This is a characteristic of being predatory).

Recommendation: Careful consideration of packet transport should be given to the transition to networks that do not use TCP.

Issue-06: **Inefficient Support of Content Delivery by TCP/IP:** The TCP/IP suite in the Internet today does not support efficient content delivery, especially video content which is expected to dominate Internet traffic in the near future. Specifically, when the E2E interconnection involves heterogeneous (e.g. radio and fixed) networks, any latency fluctuation and packet error at any part of the E2E path will degrade the end-user device's transmission throughput due to congestion control mechanism. Furthermore, slow-start mechanism means small file download (e.g. < 5 MB, which is common in webpage and video streaming sessions) often suffers from slow throughput, especially under long network latency.

Recommendation-01: If TCP is to be used, then NGP should introduce smart content handling mechanisms to reduce transmission latency through localization. For example, the access network edge may pre-fetch and/or cache the content beforehand. Such features may be realized through a dedicated network function/entity at the network edge, which further enables the option of embedding context-aware intelligence at the access network.

Recommendation-02: NGP should introduce a new internet protocol that does not employ any slow-start mechanism.

Recommendation-03: NGP should introduce new policy based networking protocols, that are able to apply flexible congestion handling techniques according to specific contexts, such as congestion avoidance or congestion control. Such new protocols may be deployed as a network function at the access network edge.

Issue-07: **Critical transmission management optimization parameters are often unnecessarily encrypted:** Bulk encryption of user traffic by a transmission protocol using E2E encryption is sometimes a customer requirement, which means it is challenging for any intermediate access network or fixed network to perform smart content management using middleware.

Recommendation-01: While providing E2E encrypted user traffic, key transmission control fields should be exposed to optimization algorithms along the E2E path.

Recommendation-02: A trusted authenticating network function should be operated at the access network edge to eliminate/minimize the challenge above by securely managing the E2E encrypted communication.

NGP should enable the access network operator to be able to embed intelligence to enable smart content management.

Issue-08: **Mismatch of Layer to Layer Bounds:** When end-user devices stream video over the Internet today, the transmission throughput is often less-than-desired, because the protocols at application, transmission and lower layers are not aware of each other's requirement or are badly mismatched with each other's bounds.

For example, in DASH streaming sessions, after user playback buffer is saturated, the user device only initiates a download once every couple of seconds (depending on the video segment length in seconds). This means the UE-server TCP connection have become idle upon the next download, and the server will perform a fresh slow-start, no matter how large the receiving window is advertised by the UE. This limits the UE throughput to the video segment size (as discussed above).

Recommendation-01: If TCP is to be used, transmission latency should be reduced to mitigate the performance impact caused by slow-start. For example, the access network edge may prefect and/or cache the content beforehand. As discussed in Issue-01, such features may be realized through a dedicated network function/entity at the network edge.

Recommendaton-02: NGP should introduce a new internet protocol without slow-start or congestion control mechanisms. Such a new protocol may operate with respect to different policies that match high-layer protocol parameters (e.g. congestion window) to the underlying layers' characteristics, such as PER and latency.

8.6.5 Applicable Use Cases (from Annex A)

8.6.5.0 Introduction

The following content enablement specific use cases are developed from the SMARTER referenced use cases in Annex A. SMARTER referenced use cases: 1, 5, 11, 12, 18, 36, 37, 38, 47 and 51 are identified as having notable performance and/or content enablement scenario dependence and/or are required in order to support next generation performance and/or content enablement enhancements.

8.6.5.1 Case 1: New Transport Protocol

The Existing transport protocol (e.g. TCP) adjust the congestion window based on the statistic of bi-directional packet drop ratio and the RTT. This kind of prediction based method usually lead to an inefficient utilization of available bandwidth and high latency when packet drops occur or the access network introduces extremely variable delay. It is now unlikely to meet the requirements of emerging service such as virtual reality which demands both high bandwidth and low latency performance. In this case, a new transport protocol architecture is required in order to meet such high performance requirements. The new transmission protocol architecture should also be able to utilize information about the component network links (including access) to realize a significantly enhanced and efficient congestion management approach(s).

8.6.5.2 Case 2: Use Case for Flexible Application Traffic Routing

When a user content request arrives, the access network should be able to flexibly route it to different sources, instead of relying on DNS server that is not context-aware. For example, the request may be routed to a nearby cache server, or it could be routed to a server that has the requested content and has the lowest latency to the user. Such routing may be subject to different policies that are based on multi-dimensional context information, e.g. user profile and network context.

8.6.5.3 Case 3: In-Network Caching

The access network edge can be capable of caching content at the network edge. It may also be capable of coordinating/managing content caches at different entities at the network edge, such as routers, etc. Such cache management can be subject to different caching policies.

Caching content at the network edge can effectively reduce the access latency at Ues, which enhances download throughput. The throughput gain is especially significant if TCP is used (as compared to the scenario where content is downloaded over the Internet from a remote server).

8.6.5.4 Case 4: Deterministic Network Reporting/Profiling

Many current networks are best effort packet delivery systems. Various types of services are applied to these current networks and contend for the available and finite bandwidth and thus the throughput and latency of each service is not guaranteed. For time-critical application, such as VR and Industry IoT, the current best effort delivery networks cannot meet the required performance of such services. In this case, future networks should either be designed with more deterministic characteristics or equipped with the ability to report current status to other layers that use their lower layers. This capability may be facilitated by quantised reporting, profiling and/or policy control.

8.6.6 Scenario Targets

Table 7 details the KPIs for improvement of this Scenario as a result of the development of NGP's.

Table 7: KPI's for Scenario - 06

KPI Name	Description	Units	Current Min Value	Current Max Value	Target Min Value	Target Max Value
ETE UL Throughput	L4 uplink throughput experienced by UE. Shall be able to support 4K and 8K 2D as a minimum.	Mbps	0	Lowest Link Rate on E2E path - Transmission Overhead	0	Lowest Link Rate on E2E Path (150 Mbit/s for 8K, 2D)
ETE DL Throughput	L4 uplink throughput experienced by UE. Shall be able to support 4K and 8K 2D as a minimum.	Mbps	0	Lowest Link Rate on E2E path - Transmission Overhead	0	Lowest Link Rate on E2E Path (150 Mbit/s for 8K, 2D)
ETE UL Pk Latency	ETE user uplink packet latency at L4 level.	ms	10 ms	2 s	1 ms	Propagation delay on E2E path
ETE DL Pk Latency	ETE user downlink packet latency at L4 level.	ms	10 ms	2 s	1 ms	Propagation delay on E2E path
ETE UL PER	ETE packet error rate at L4 level.	%	0,1	0,4	-	0,001
ETE DL PER	ETE packet error rate at L4 level.	%	0,1	0,4	-	0,001
AccessDelay	The time duration it takes between - UE sends a request, and - UE receives the first content payload packet.	Seconds	10 ms	Any	1 ms	Propagation delay on E2E path
RebufFreq	The frequency of rebuffering events during a video streaming session.	Times	0	Any	0	0
RebufDur	The total duration of rebuffering events during a video streaming session.	Seconds	0	Any	0	0
Jitter	Jitter refers to the rate of change of latency. The lower the measure of jitter the more stable a connection is and it is important to gamers, VoIP users and other interactive applications.	ms	0,5 ms	1,6 ms	-	0,5 ms

8.7 Network Virtualisation

8.7.0 Introduction

This clause addresses virtualisation scenarios to be considered for NGP. The scenarios are described so as to identify key issues with:

i) complexity;

ii) flexibility; and

iii) ease of adoption noted when seeking to adopt network virtualisation that includes both current core, and an access networks (e.g. when working with internetworking systems such as experienced with Wi-Fi™, Cellular, mm-Wave access technologies connecting with fixed infrastructure and the internet.

There are three aspects of network virtualisation that are considered here:

i) Network Virtualisation (NV) which is an over-arching network virtualisation approach for next generation networks (NGN) that is independent of infrastructure and may include such logical network virtualisation entities as Network Virtualisation: Orchestrators, Controllers and Agents. Current implementations include elements of SDN and NFV.

ii) Software Defined Networking (SDN) where L1/L2 hardware transmission components are separated from logical flow control which is managed in a soft manner using SDN-VIM/Controllers, using southbound protocols (SB-P) to SDN-VS interface.

iii) Network Function Virtualisation (NFV) where traditional and new functions are implemented entirely as software functions or Virtual Network Functions (VNF). This branch of virtualisation has been extensively standardized in the ETSI ISG NFV: Requirements: ETSI GS MEC 001 [19], Services: ETSI GS MEC-IEG 004 [21], Architecture: ETSI GS NFV 002 [8] and MANO: ETSI GS NFV-MAN [17].

Issues related to both SDN and NFV should be considered with respect to NGP, and solutions progressed to improve network virtualisation drivers (reduced complexity, improved efficiency and simpler ease of adoption). Addressing these drivers for better network virtualisation enables operators to gain the full value of network virtualisation as follows:

1) Rapid service deployment of VNFs and NSs onto COTS blade based hardware
2) Simple network re-configuration
3) Simpler upgrade cycle
4) Reduced OPEX
5) Virtualisation
6) More efficient utilization of resources
7) Clean management of resource/isolation of resource and traffic (data) on shared network
8) Abstraction of physical infrastructure to maximize decoupling and programmability
9) Easier to realize (in-Network) Self Organizing Network functionality (extending SON from RAN to ETE)
10) Easier to provide Multi-Vendor support

8.7.1 Model Architecture

The scenarios of virtualisation are covered in broad categories, as follows:

1) Virtualised distributed service model
2) Network slicing (partitioning)
3) Multi-tenancy in fixed and mobile networks
4) Virtualisation of radio, core and transmission resources

Figure 37 shows a high level description of current mobile network orchestration, and the network domain segments.

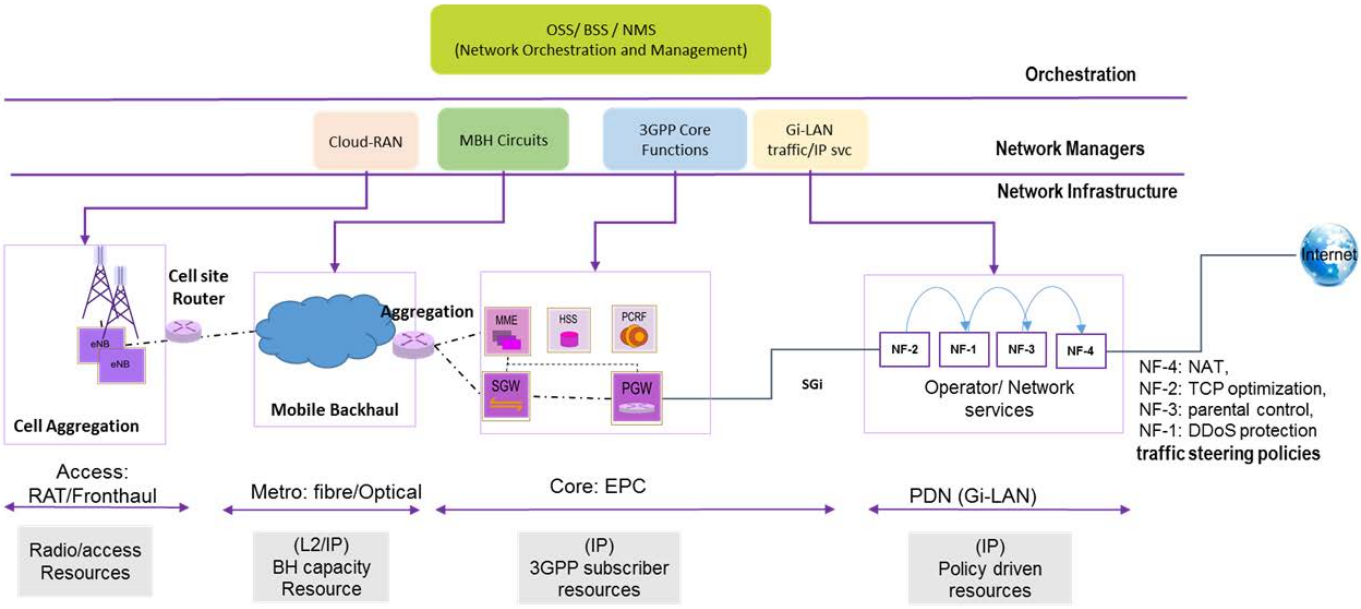


Figure 37: Management and Orchestration of Mobile networks

A mobile network consists of different network segments in administration domains that are managed separately by network controllers. These segments depict some degree of virtualisation in the following manner:

- C-RAN (Cloud or Centralized-Radio Access Network): C-RAN virtualises the radio Access network (RAN) by connecting many Remote Radio Units (RRU) together over 'fronthaul' towards a pooled Baseband Unit or BBU which operates as the anchor point of the S1 interface towards the core network. The BBU is the logical eNB interconnection point to the connected pooled RRUs.

NOTE: Fronthaul is the transmission between the RRU and the BBU and is usually deployed as Dark-Fibre for the current 3GPP specified C-RAN and operates various forms of CPRI.

  In this centralized approach, C-RAN virtualises the RAN and enables better Layer2 coordination between the Radio units in terms of such RAN optimization features as: pooled resource sharing, CoMP, Massive MiMO, multi-point transmission optimizations and inter-cell interference coordination ICIC.

- H-RAN Distributed RAN (D-RAN) or traditional RAN technology is also very common today but operates a lumped element single physical entity build type of Base Station. These types of base station are widely deployed in many types of environment for Cellular LTE and are difficult to operate some of the higher layer levels of inter-BS coordination for RAN optimization. However, LTE stage C-RANs require the deployment of Dark-Fibre for their fronthaul, which is very expensive for some suburban and most rural area types. As such, the RAN study report in ETSI TR 138 913 [i.25] on new radio for next generation cellular networks is introducing the concept of a Hybrid RAN or (H-RANs) that combines the best of C-RAN and D-RAN. Whilst ETSI TR 138 913 [i.25] acknowledges that C-RAN is still a good option for Dense Urban environments and that D-RAN is the only cost effective option for some rural deployments they are seeking to define a more generally optimal approach for pooled RAN equipment in their H-RAN studies to enable inter-BS coordination and RAN optimization, whilst operating with various cost-effective fronthaul options.

  H-RAN will also be operated with RRH and BBU physical equipment separated by fronthaul, but the break in the protocol stacks between these entities is likely to be at a different point than for C-RAN today so as to avoid the need for Dark-Fibre interconnection.

  These H-RANs will also need Network Virtualisation.

- Mobile Backhaul network: Mobile backhaul is the transmission that connects the eNB (D-RAN or C-RAN) to the cellular core network. D-RANs may be directly connected to an aggregation point CPE or PE router or they are often deployed with a co-located CPE router if there are multiple Base Stations at the same site. A 'cell site' router drives the traffic through mobile backhaul to reach the mobile network core and from this point onwards all transport is IP. These connections may be leased lines or self-owned fibre, or metro networks.

The mobile backhaul situation is similar to C-RANs but they often include routing capabilities themselves or connect to the co-located CPE or PE router at the same datacentre as the BBU itself.

- Evolved packet core (EPC): this functional group can be virtualised with all of the 3GPP functional elements HSS, MME, P-GW, S-GW and PCRF implemented as VNFs The EPC is evolved at 3GPP, Rel-14 stage into a Control and User Plane Separation (CUPS) evolution of its gateways as control parts S-GWc, P-GWc and user parts S-GWu and P-GWu for greater flexibility of use and N to N mapping between elements being enabled which is a good fit for Network Virtualisation.

- Packet Data Network where L4-L7 network services are virtualised network functions (VNFs) and service chains of these (see SDN and NFV refs in this clause earlier) are implemented to provide Sgi-LAN services before traffic can be sent to the Internet.

This model has the following issues:

- It is currently very centralized and all traffic converges to EPC and more so to the PGW often only being one logical PGW per network in current deployments.

- With many access networks available, switching from one access network to another involves many signalling messages involving different attach/association, authentication, addressing, customer profile management, routing and charging functions.

  The users IP address needs to change:

  i) when re-attaching to a cellular network; and

  ii) when switching from one access network to another. Also, the use of APNs for different access to different IP networks often necessitates that the user needs to gain a different IP addresses from the cellular/access network for access to each of the connected IP PDN networks that they wish to connect to. (e.g. Cellular Intranet, Corporate VPN and Cellular IMS network). As the UE, mobile nodes and IP network elements in EPS grow - the reachability and interconnection of these require sophisticated, lightweight and efficient IP routing.

- Both Mobile and fixed network are expected to provide similar services (video, broadband, cloud computing, enterprise class services) with comparable usage, user experience, security and accessibility. This necessitates the need for a more flexible packet routing approach for the mobile core.

- Segments in the EPS (e.g. C-RAN, vEPC and cloud based Gi-LAN) are independently virtualised and managed. First a much simpler, and an integrated orchestration and management is required to streamline coordination across these segments together. Secondly, instead of API loaded models, a more agile resource discovery and distribution protocol is needed.

Evolved Cloud based service model

An evolved model architecture is presented in detail in Figure 38. The main idea here is to push computing to the cloud:

a) in proximity to the UE to provide ultra-low latency and high-bandwidth for critical applications;

b) in public cloud for compute, and storage intensive applications.

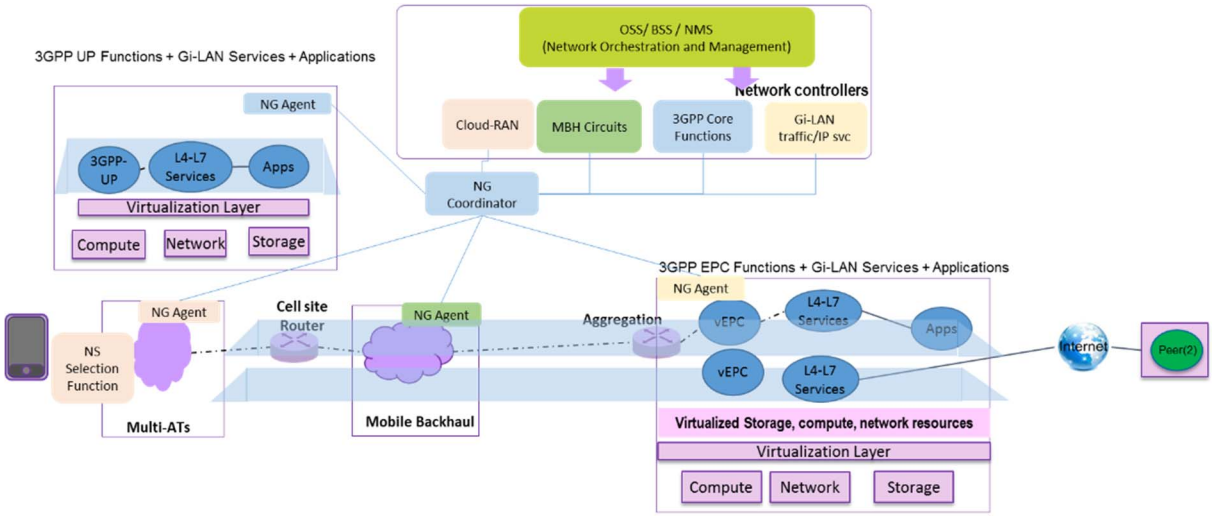


Figure 38: Virtualised Service and Application Distribution (Evolved Model-1)

In this evolved model, a cloud based EPS/Next Generation Cellular approach is envisaged as follows:

a) The ETSI model of NFV is fully adopted to include SDN/NFV VIMs and Orchestrator. However, this model is further evolved to include Agents and Next Generation Coordinator (NGC) that is introduced to provide granular sub-network management and orchestration in sync with the currently proposed NFV architecture and framework, as follows:

- This model presents combined concepts from the ETSI, ISG NFV MANO definition of the orchestration platform [17], the ETSI, Mobile Edge Computing ISG MEC Architecture [20], Requirements [19] and Service Requirements [21] and Network slicing as detailed in the options in 3GPP TR 23.799 [i.2]. In 5G, cloud based service models will be prevalent and isolation of traffic across each service or slice will be required through multi-tenancy (i.e. infrastructure sharing).

b) The NG coordinator works with each of the following domains managed by the orchestrator:

1) Access Network: RAN

2) Access Network: Fixed

3) Transmission

4) Core Network

5) Content Networks (represented as data centres)

To further clarify, transmission refers to a separate network (say MBH) that interconnects RAN with the core, Within the scope of a controller the control of transmission nodes implies intermediate nodes such as routers, switches or optical nodes.

c) Several dimensions of Multi-tenancy are possible in this model, i.e.:

1) a virtual or private enterprise network that uses multiple access technologies (3GPP and non-3GPP);

2) service aware network with specified resource allocations; and

3) an isolated independently managed mobile virtual network that could offer different services besides broadband (video, gaming, location based information, etc.).

d) Since EPC is virtualised, some network functions of EPC are moved closer to the subscriber (UE) access or collocated within the RAN this approach is commonly called a 'flattened' cellular architecture.

e) Virtualised network services (Sgi-LAN style) can be instantiated on access closest to the subscriber on-demand, based on application attributes (lowers round trip latency).

f) Each multi-tenant instance maybe treated as a virtual network, which consists of pre-defined resources (L4-L7 services, applications, radio and capacity).

g) Multi-tenancy is realized through mobile network infrastructure independence with the aid of 'NG coordinator' and 'NG agent' that together distribute and discover virtual networks. These two entities can also be extended to provide agile and flexible discovery, sharing and coordination of resources from the infrastructure.

h) Integration of virtualised RAN is not included in the slice, which is same as 3GPP TR 23.799 [i.2] and keeps network slice IP based.

Advantages of this Model

a) Applications are distributed, location aware, and instantiated on demand in proximity. This is similar to MEC (Mobile Edge Computing) concept but deployed in virtual instance of mobile network for infrastructure economy; Thus, providing higher cost-efficiency through better resource utilization, reduced energy consumption and network isolation (Optimal path between subscriber and applications eliminate need to go through backhaul).

b) The model has a network slice selection function in RAN edge (similar to that proposed in 3GPP TR 23.799 [i.2]) where radio to packet conversion is operated (which reduces latency).

c) Based on the above points, the model understands and supports three identifiers:

   i) a subscriber;

   ii) a virtual network; and

   iii) a service.

   A subscriber can subscribe to multiple virtual networks and a single or many service(s) within that virtual network.

8.7.2 Scenario Description

8.7.2.1 Scenario #1: Network Virtualisation in EPS

This scenario is based on network slicing concept from 3GPP TR 23.799 [i.2] - *"An ability to create networks customized to provide optimized solutions for different market scenarios which demands diverse requirements, e.g. in the areas of functionality, performance and isolation"*:

a) A mobile network infrastructure operator offers its 'network resources' to various virtual mobile service operators (vMSO) as a network slice. A UE can use services from different vMSOs at the same time for example:

   a) a corporate virtual or private network;

   b) vehicle to infrastructure (V2X) network;

   c) broadband.

b) It is envisioned that distributed data centres are available in the mobile network to offer 'cloud computing' platforms (PaaS, SaaS) for service hosting in proximity of the users. The core mobile network operator provides network resources to a vMSO as a network slice, the traffic isolation is achieved through network virtualisation. A vMSO is then able to flexibly orchestrate and interconnect the services with in its slice without over-stepping on resources used by other slices. For example, MEC based use cases bring services closer to the user and a private cloud connects employees with localized enterprise applications.

   In both cases, the data, content, compute and storage are hosted in a cloud (or data centre) anywhere in the mobile network (access, aggregation, EPC, and Internet). The most important task in NV is to determine the closest location of different types of services.

c) Minimize configuration and Orchestration overheads through autonomic networking: The goal is to enable infrastructure independent coordination of resources. The current orchestration methods are static - services and service chains are templated; repository is built through slow management techniques - configuration, SNMP, REST APIs, etc. Therefore, NG protocols for virtual routing will facilitate auto discovery of slices, all nodes in the slice, and their dynamic addition, removal, and elastic scale out scheme.

d) A UE connects to different network slices through its user context. The network attachment request is serviced by a Network Slicing Selector Function (NSSF). The NSSF learns about different slices through new virtual routing protocol between NG coordinator and NG agent. Using the existing GTP based model, multiple GTP tunnel ids per UE will be required to allocate QoS and the default bearer channel for each slice. Otherwise, with a single tunnel id it is not straightforward to distinguish UE's slice context and per session state-full mapping will be required on a UE.

e) On the network operation side, vMSOs have isolated virtualised orchestration systems to independently manage their own network slices. For example, create their own custom service templates and multi-instance chains of virtual network functions.

8.7.2.2 Scenario #2: Virtualised RAN

In the context of RAN, enhancement on can be applied to different radio access aspects:

a) spectrum enhancement on, which allows multiple network operators to share the same spectrum for a more efficient utilization;

b) hardware and network sharing, which reduces over-provisioning especially for small cells with the aim to reduce both OPEX and CAPEX;

c) multi-RAT enhancement on, which simplifies the management of different RATs, where each of them is dedicated to support a specific service or offer a different QoS;

d) computing resources enhancement on, which is used to share the computational resources available at a central processing center (a.k.a., BBU pool) among multiple BSs. In the applicable literature, this architecture is referred to as C-RAN, where the "C" stands for cloud, central, collaborative, cooperative or clean. A general and simple description of this architecture is depicted in Figure 39.

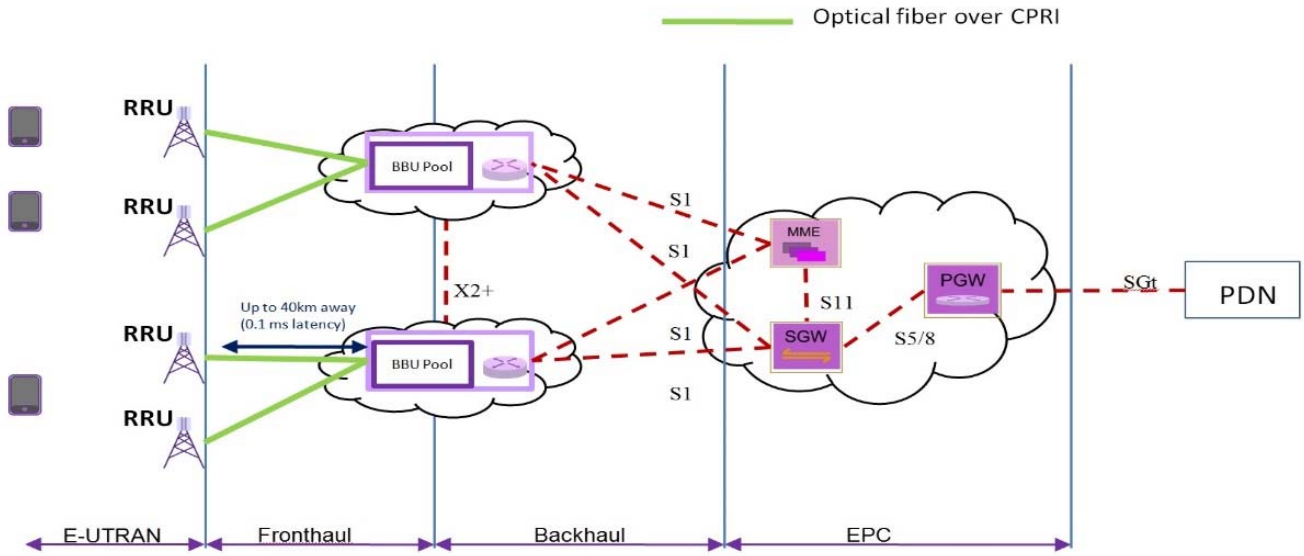


Figure 39: C-RAN LTE mobile network architecture

Virtualisation of the radio access technologies allows:

1) To cope with temporal and spatial traffic fluctuations in mobile networks.

2) A better scheduling of the computational effort on critical operations in difficult channel conditions.

3) Partially centralized execution of baseband functionalities depending on the actual needs as well as network characteristics and backhaul capabilities, throughout flexible split of (signal processing and resources management) functionalities at the PHY, and MAC layer RAN as a Service (RaaS).

In a virtualised RAN system, radio and data-processing resources should be managed jointly: the allocation of radio resources should occur by accounting for the channel condition and the QoS requirements, but also by accounting for computational resources demand. This tasks should be predictive (the system should be able to estimate the computational budgets, and allocate the RAN resources accordingly), and should involve auto discovery of available resources either within a BBU pool or among BBU pools (for instance throughout the X2+ interface). Specific use cases are described in clause 8.7.4.

RAN improvements are heavily reliant on intercommunication between site based radio cell equipment and mobile UE devices. To enhancement RAN performance, RAN optimization techniques such as radio relay, coordinated multipoint (CoMP), inter-cell interference coordination (ICIC), and distributed antenna systems (DAS)/MIMO and Massive MIMO are in continual evolution at cellular standards.

Many of these techniques have very tight time constraints (e.g. CoMP requires 0,5 - 3 μs timing accuracy for LTE). Such decisions are best made at the network edge and require little or no information from the wider network. In this context, virtualised RAN can be used to construct a software-defined abstraction of the physical radio units (including non-3GPP) in a local geographical area, as a virtual macro base station (i.e. SoftRAN, V-Cell), comprising a central controller and radio elements, which appear at the EPC as a sole physical entity. The core of the virtualised radio station can then perform three main functions:

i) mobility anchoring;

ii) mobility management;

iii) Self-Organizing Networking (SON) in the RAN (optimization, healing, and organization) see 3GPP SON Requirements ETSI TS 132 500 [24].

8.7.3 Applicable Issues

Issue-01: **Virtualisation (Multi-Tenancy in Mobile Networks)** will be required to drive new service models by sharing physical infrastructure; it helps efficient use of resources. The mobile networks lack agility to elastically align resources as well as isolating a group of subscribers through virtual networks. In future, mobile network will serve as an infrastructure, with multiple virtual network operators or tenants using the resources.

Recommendation: As the most parts of mobile network become IP (or NGP), an end-to-end network isolation should be supported. The tunnel IDs themselves are not sufficient and additional overlays will be needed. It is recommended that encapsulation starts from the access, so that a UE device is agnostic to any encapsulations.

Issue-02: **Network Slicing:** According to 3GPP TR 23.799 [i.2], slicing is limited on a specific service profile concept (e.g. broadband, V2X, etc.). Another way to slice the network is allocation of resources all the way from virtualised access technologies to EPC, and create an instance of virtual networks (including MBH and RAN)**.** With predetermined QoS in each segment, this allows to estimate latency more accurately. In order to flexibly utilize network abstraction a slice should be a top level virtual network within which several other services can be organized/offered.

Recommendation: NGP should support the following logical virtual networking features realized through the use of network virtualisation (NFV/SDN/MANO):

- Network Slice: a virtual network realized through network virtualisation.
- Network Service: a virtual network function graph realized through network virtualisation.
- Logical Subscriber Grouping: a logical grouping of one or more Network slice or network service subscribers operating devices that are able to connect with/access that slice or service.

In this manner, a unique session is a tuple {Network Slice, Network Service, Subscriber Group(profile)} and granular user context, traffic QoS filtering may be applied to each subscriber group virtualisation.

Network virtualisation may be applied using traditional GTP tunnelling, but the NGP team recommends adoption of more native networking techniques based on internetwork routed technology, as more efficient.

Issue-03: **Scale:** As the mobile network infrastructure becomes virtualised, the number of VNFs and mobile subscribers/mobile devices s will increase significantly over time with their own addresses, isolation and reachability requirements.

Recommendation-01: NGP should provide simple routing schemes to support network virtualisation that do not require large IP address ranges or mappings.

Recommendation-02: NGP should widely adopt datacentre virtualisation techniques that minimize overheads at the infrastructure management level.

Issue-04: **Security:** As the mobile infrastructure is shared, multiple instances of similar looking 'network services' will be common in the future and they should remain unaware of each other's presence. The trusted domain concept needs to be redefined with not only stronger encryption and authentication but also strict pre-deployment verification checks and certificate checks before a network function is added in to service chain. The NFV SEC (security team) have identified threats [18], [22] and [23] that maliciously damage the operation of the intended SDN/NFV/MANO framework deploying VNFs and co-hosted VNFs running on the same connected virtual infrastructure and Maliciously use, distribute, adjust or delete the information that is passed by VNFs.

Recommendation: NGP should improve the security of the underlying transport that supports the network virtualisation distributed control framework. With underlying security, many control functions in NFV systems can be simplified to a great extent. The main issue to be addressed by NGP is to provide inherent Registration/Attach/Association with a network for transmission of Infrastructure critical transmission systems.

Issue-05: **Management and Orchestration** has evolved into a complex centralized multi-layer service architecture that includes repository catalogues, virtual infrastructure manager, service chains and network controllers. A service provider builds an application logic to co-ordinate resources across multiple network segments. While all these are necessary components, the method of allocating and managing resources is operator driven (manual), complex and centralized.

Recommendation: NGP should identify a management protocol that is a balance of centralized control and autonomic control in order to reduce dependency on application logic (thus MANO platform). This capability is envisaged to include protocol agents that make decisions closer to where resources are enabling agent to agent query/subscription capabilities to be realized, about resource trading based on:

1) pre-defined policy constraints and contracts configured by MANO's application-logic and SLAs;

2) current availability known to the resource agents.

Issue-06: **Programmability efficiency for NFV/SDN:** The transmission requirements for control information are different for the data traffic. In NFV solutions, often southbound interfaces (SBI) are used to program network infrastructure. Southbound protocol (SB-P) may run on TCP/IP because it is important to acknowledge critical changes (flow programming, switch port configuration) in the network. Whereas, often such control messages do not require high throughput and congestion control (CC) mechanisms. In order to achieve the benefits of network abstraction through SDN/NFV the transmission of such control messages require a much lighter yet reliable data transmission.

Recommendation-01: NGP should include the provision of an enhanced transmission protocol that is suitable for virtualisation that scales from full CC to no CC depending on the status of its lower layers.

Recommendation-02: The enhanced transmission protocol of Recommendation-01 should be customizable for both control and data traffic requirements of reliability and bandwidth.

Recommendation-03: Optionally, within the scope of network virtualization, NGP should investigate methods to achieve reliable data transmission for virtualization support without congestion control overheads. An example of such an approach could be UDP plus virtualization tailored ARQ control, as a first step.

Issue-07: **Resource management in a Soft Cell Environment:** Next Radio Soft-RAN's (C-RAN and H-RAN) provide a mobile device with simultaneous logical connections to a heterogeneous layered cellular access framework of cells using a range of non-overlapping frequency bands in a highly densified network with the aim to provide higher capacity, while maintaining mobility and continuous connectivity. All cells in this evolved RAN environment share the same control channels. Control and user plane are decoupled, and the handover takes place at a cell level, and there is no need to redirect traffic between neighbouring nodes as handover is anticipated to be soft or very fast. In this environment the mobile device needs to be able to support multiple transport channels, and MAC entities, along with scheduling of the activities related to the master and slave cells/carriers. The Mobile device needs to also be able to promptly discover new 'in-range' smaller cells, while maintaining connection with macro-cell (3GPP defines this capability as Dual Connectivity (DC) - separation of CP and UP in a multi-cellular, multi-layer cellular environment). For a fast, and energy efficient discovery procedure for small cells, a very tight synchronization is required. The spectrum allocation is static.

Recommendation-01: A coordinator or centralized unit (i.e. C-RAN) may be used to perform a more efficient and dynamic spectrum allocation with a higher degree of re-use on a demand and intelligent context basis, and a more energy efficient management of the small cells (small cells that are not used can be simply switched off).

Recommendation-02: Handover within a macro-cell may be done in a proactive manner with limited or no signalling between the UE and its camped on small cells (this relaxes the synchronization requirements and related time constraints) using the information collected by the coordinator using primarily or solely the macro-cell. Interference can be limited throughout centralized frequency reuse and beamforming coordination. In order to sustain the scalability of this architecture, virtualisation should be applied to abstract the whole macro-cell comprising of the small cells. By virtualising this portion of the network, the full benefits of virtualised RAN can be gained.

Issue-08: **NATs processing load and delay:** There is a very large use of NATs in virtualised networks since it helps reuse efficiently IP addresses, provides an extra layer of security, and it hides the internal network topology of the virtualised network. However, this has several drawbacks:

i) NAT is unable to support some applications (peer-to-peer applications, i.e. voice over IP), where the initiator lies outside the subnet unless a fragile, complex and tedious procedure is implemented;

ii) NAT does not have a clear and uniform response to fragmentation, as it normally relies on the TCP/UDP header for translation;

iii) makes the management of the network more complex (either computationally or in terms of memory), since the NAT devices add state to a specific location in the network with the implication of causing delays related to the NAT translations and other operations (i.e. check sum);

iv) it is not a practical solution for large numbers of internal hosts all talking at the same time to the outside world;

v) it increases the likelihood of errors in addressing.

Recommendation: NFV, and independent management of mobile virtual networks are good tools in the context of the proposed architecture to reduce NAT operations and therefore reduce NAT limitations for certain applications and some specific services. NGP should avoid the need for NAT, throughout the use of a more efficient solution. This new solution should be simple, agile, scalable and application independent.

Issue-09: **Net Neutrality Legislation and Network slice adoption:** The proposals for slicing are subject to restrictive nature of net neutrality. The "5G Manifesto for timely deployment of 5G" [i.24] in Europe has raised the legislative restrictions of net neutrality do not allow adapting in real time to changes in end-user/application and traffic demand due to the equal and non-discriminatory treatment of the traffic. The issue is mentioned as informational that the group is aware of such developments, however, it is not a technical problem that NGP needs to resolve.

Recommendation: NGP should bear in mind any implied constraints of Net Neutrality legislation.

8.7.4 Applicable Use Cases

8.7.4.1 Case 1: Network Slicing

The NG network slice use case in Figure 40 shows many virtualised components - the core network has multiple instances. Network slicing is expressed through network virtualisation to isolate resources and traffic. The network nodes (or functions) in each instance of EPC or elsewhere in network have an IP address. In this example, a user may be interested in three (3) different services that may belong to same slice. A slice may represent an enterprise cloud network that has applications distributed across:

a) the Internet;

b) co-located at EPC; or

c) in user proximity (RAN).

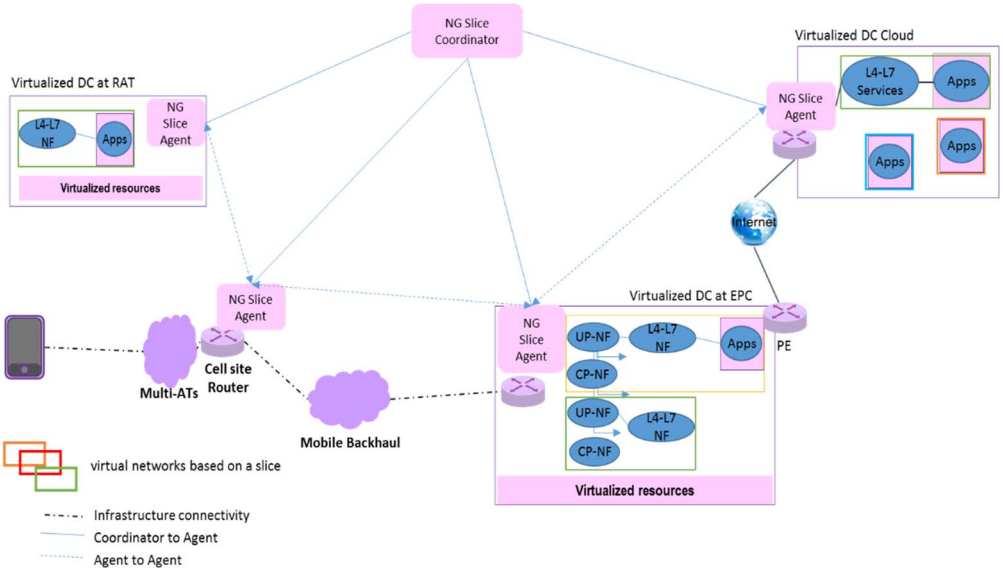


Figure 40: NG Slicing

This concept enables infrastructure independence and auto-discovery through two network elements:

i) Next generation Agents (NGA); and

ii) Next Generation coordinators (NGC):

a) An NGC coordinates NGA (tunnel endpoints) discover different applications in the cloud.

b) A DC infrastructure operator configures a well-known slice through OSS into NG Agent.

c) All NGA accordingly register with coordinator the slice instances they host and discover other NGA with same slice from the coordinator.

d) The virtual network is created on demand when NGA exchange reachability information of all the network, service or application nodes among themselves on per slice basis along with their location in the infrastructure network.

e) From originating NGA, any user plane data exchange can be encapsulated with slice-id (or virtual network id) for communication between 2 nodes in a slice. The de-encapsulation is performed by the destination NGA. The process of Encapsulation and De-encapsulation is termed here GVE (Generic Virtual Encapsulation).

Only a virtual network representation is discussed, there is no impact on how service chaining is done and OSS instances are assumed to define the VNFs and VNF graphs. The scenario allows a slice to be a virtual network or multiple virtual networks created within a slice. Although GVE is shown to carry virtual network information, this may raise a question about the increased payload size, this can be correlated to using VXLAN type encapsulations in the data centre to provide network virtualisation. In fact, GVE here can be VXLAN.

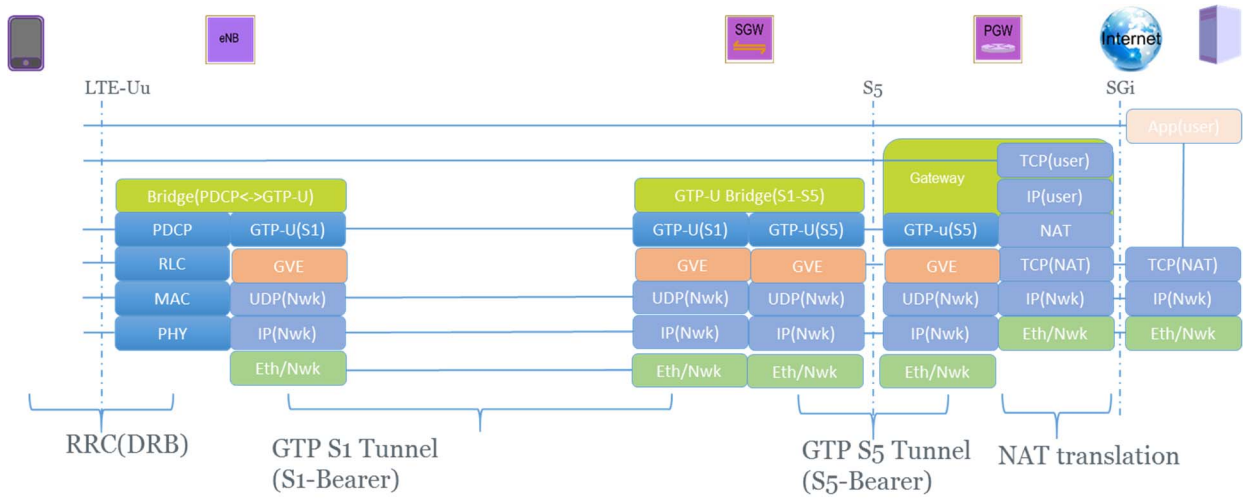


Figure 41: Network Slice instance in User plane

NOTE 1: In a GTP-free NGP stack, which is assumed ID aware, it may be beneficial to have NGP over NGP tunnels, in this case the ID in the outer header will represent a slice/virtual network. Several IP replacement protocols will be considered for access networks by NGP and GVE is anticipated to operate over these using a Slice binding and/or Encapsulation and/or setup between. The envisaged position of the GVE in the existing cellular LTE is illustrated in Figure 41.

NOTE 2: In this use case, OSS specifies a slice and pushes the Configuration Management (CM) information to 'build' the slice to all NGA t affected. The NGC facilitates auto-discovery of VNFs. It is assumed that the mobile device/subscriber user context is extended to carry service context and slice ID. A Network Slice Selection Function (NSSF) maintains the association of the context and slice.

8.7.4.2 Case 2: Network Slicing: With Simultaneous access to different instances of Virtualised core

A user surrounded by multiple access networks and many service operators in mobile network will have to choose to select the best suited service offering based on context, cost and requirement balance. In addition, the market context will also vary, for example, a service for ad-hoc vehicle network logic will be concerned with low latency, low capacity and instant update of information, which is a different context-set from the requirements for a predominately content delivery algorithm set that could absorb some delays but require higher bandwidth along with in network video encoding functions.

Thus, a mobile subscriber/device should be able to simultaneously register with different service providers. A service context is therefore mandatory to identify what service/context is being requested.

8.7.4.3 Case 3: MEC and Network Virtualisation

Consider a use case where a UE requests access to popular high definition video content. Since the content is in high demand accessed by many users in the mobile network, it is cached in the nearby data center connected to RAN. This content is replicated across different RANs. Using an NGA. NGC, the NSSF at the RAN will discover and determine the closest available server instance of the requested content. ETSI MEC has also considered interesting mobility scenarios with MEC servers, i.e. when a UE moves while accessing the content. This feature suggests that user get a seamless experience whilst watching video on the move.

8.7.4.4 Case 4: Cloud interconnect (Mobile/Fixed networks)

This use case is an interconnection of clouds that may be connected through 5G access, LTE (3GPP) or fixed access. In one of our previous cases, the association of slice to a virtual network identifier was discussed.

Consider an enterprise's campus site in mobile network connecting to the main site hosted as a virtualised data center. The scope of a slice is defined within the mobile network; it will terminate in EPC before exiting to the Internet. This implies that in order to support virtual network or VPN, this packet has to be re-encapsulated before exiting mobile core.

As the cloud based enterprises and the endpoints in mobile network grow, the demand for IP address reachability and secure connectivity can make network management complex. Instead of employing IGP and MPBGP VPNs that are more suited for less-frequently changing network infrastructures, a generic and lightweight virtual network discovery, and route distribution feature can be affected using the same scheme as outlined earlier in this scenario section (using NGC and NGA to execute SON algorithms for route optimization across the virtualised infrastructure.

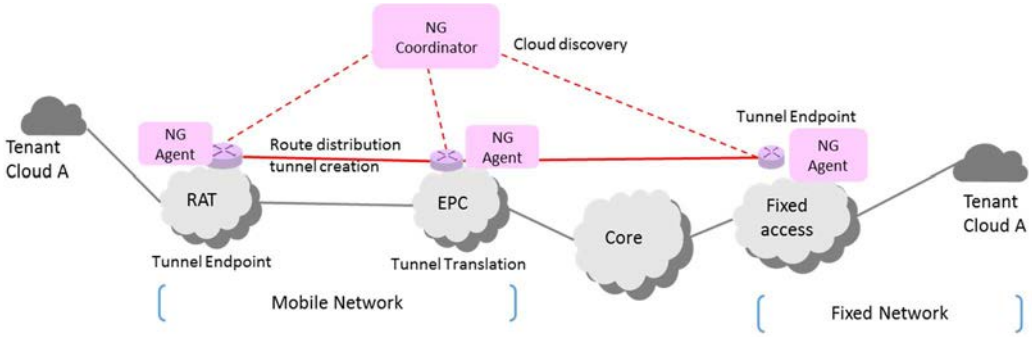


Figure 42: Cloud based Interconnection

NG cloud interconnect

- In Figure 42, the same illustrates the same approach as for clause 8.7.4.1 earlier illustrated, but the concept is extended to interconnect by auto-discovery enterprise clouds on a potentially global basis.
- Auto-discovery and route distribution per cloud is affected in a similar manner to the earlier slicing use case for the EPC on behalf of campus networks in the EPS.
- The association of campus cloud route distribution is learnt at the EPC NGA by exchanging routes with the remote NGA (in fixed access cloud). In this case, after the GTP header is stripped, it goes through an Sgi-LAN NF-graph and determines that it needs to set up encapsulation of the packet towards the remote NGA.
- In another case, which operates an NGP evolve, GTP-free user plane, the NGA could directly send GVE encapsulated packets towards the remote NGA.

This allows a converged virtual network solution for cloud centric networks to be distributed anywhere. In contrast to a traditional BGP-VPN, a lightweight control plane signalling system is provided to support global reachability that provides more flexible IP address management when VNFs are virtualised in a mobile network.

The same principles are expected to work even better for NG architectural proposals that remove GTP-U.

8.7.4.5 Case 5: C-RAN Enhanced Computational Flexibility

This scenario is in reference to C-RAN/H-RAN architecture of 3GPP, or other architectures that are characterized by local provision of computational resources (BBU pool) that are jointly shared among a pool of RATs. If the computational resources are not sufficient to perform baseband processing at a given time, packets are dropped and communication is interrupted regardless of the quality of the signal. This event is called in the literature computational outage. In this scenario, two cases are considered, as follows:

a) Lack of computational resources within a BBU pool: computational requirements are correlated to the baseband processing (mainly, error correction decoders). The resources available (i.e. modulation and coding scheme) can be chosen by being aware of the computational limitation of the system with the aim to improve the overall performance of the network by minimizing computational outage.

b) Cooperation among BBU pools enforced by auto-discovery of the resource available among BBU pools in emergency situations: During critical events (earthquake, tsunami, etc., which can lead to a high densified scenario, or damages to the BBU pool) minimal connectivity needs to be still guaranteed for a large amount of users, while computational resources are locally insufficient (at the BBU). In this case, a centralized conversational scheme can be used to discover underutilized BBU pool and share them among BBU pools.

Limitation: BBU pools are self-managed. No co-ordination of resource sharing between them.

Recommendation: A new network function that is an integrated C-RAN/H-RAN coordinator can be introduced, which gathers light-weight information and updates on the health and status of the BBU (for instance, they can signal their interest to "borrow" computational resources from another BBU, because some nodes in the cluster have failed or a great densification has drastically occurred) through a cloud-RAN coordinator (i.e. it can be a new network function in PGW or a manager in MANO):

a) Maintains a dynamic state of BBU pools and makes decisions on how BBU pools need to be connected throughout the X2+ interfaces providing networks routes automatically. For instance, once a BBU requires extra computational effort, which cannot be managed internally by the BBU pool, it can be timely connected by the cloud RAN coordinator through a specific X2+ interface to another BBU pool, which is underutilized to maximize the resources available by still trying to meet time constraints.

b) A cooperating communication protocol shall allow automatic announcement, allocation and release of resources on per logical network basis to create self-organizing RANs.

8.7.4.6 Case 6: Heterogeneity of RAT

As future networks are expected to be more highly densified, it is also predictable that they will also have a much higher degree of heterogeneity with a coexistence of different radio access technologies that include both macro cells and smaller cells, such as micro/pico BS, low power RRH and femto cells.

In this context, the H-RAN/C-RAN architecture can be utilized with heterogeneous networks (Het-Nets) in order to take advantage of the full benefits of both technologies (a.k.a. H-CRAN). If a front-haul exists between the BBU pools, and all heterogeneous access nodes, the BBU pool can serve as a coordinator to provide centralized compute and processing for control plane functions with the aim to allow them to coexist and cooperate. In this context, throughout a centralized pool, radio resources can be more efficiently allocated or moved around with in and across network slicing. There are two (2) options:

i) software-defined mobile network control, which is commonly used today;

ii) self-organizing communication protocol as mentioned above.

8.7.4.7 Case 7: Performance Enhancement of Low-power RRU

H-RAN/C-RAN can be used to preserve energy consumption of the low-power access nodes, and also to guarantee service when the processing cannot be done locally at the access node, and this does not have (or it is not practical to have) a front-haul that connects it to the BBU pool. The high power RRHs can be used (consistent with the technology) as an intermediary between the low-power RRHs and the BBU pool (when this is possible). For instance:

1) When a low latency and high throughput front-haul between some low-power RRHs and the BBU pool is not (cannot be) established, and there are some underutilized and available higher-power RRH, these can serve as a relay to process the signals in the BBU pool with the aim to preserve energy consumption, and minimize computational outage.

2) Co-ordination among low-power RRHs can be established in the BBU pool by using the high power RRHs. These RRH can serve as a relay between BBU pool and low-power RRHs only for their signaling, while the co-ordination is done in the BBU pool.

8.7.5 Scenario Targets

Table 8 details the KPIs proposed for improvement of this Scenario as a result of the development of the ETSI ISG NGP.

Table 8: KPI's for Scenario - 07

KPI Name	Description	Measured feature	Current behaviour	Target Value/behaviour
Nv_ngp_slice_01	Network Slice: creation of multi instance Virtualised Infrastructure	Service	None	Creation of multiple virtualised EPS networks co-exist. Network behaviour more secure, Meets or exceeds performance of REL12 LTE Networks.
Nv_ngp_slice_02	Network Slice (UE): simultaneous access to different instances of Virtualised core	Service	None	No degradation of service guaranteed by operator. Meets performance of service as if offered in physical EPS.
Nv_ngp_slice_03	MEC with Network Virtualisation	Service	Without NV	No service degradation. Nice to have - ability to auto-discover nearest MEC server.
Nv_ngp_slice_04	Cloud interconnect (Mobile/Fixed networks)	Service	Static address mgt and NATs	Eliminate address translations.
Nv_ngp_rat_01	CRAN enhanced computational flexibility		Isolated radio resource control	A zero config, self-organizing resource coordinator. Dynamic resource allocation.
Nv_ngp_rat_02	Heterogeneity of RAT	Service	Isolated BBU pools	Same as above, dynamic radio resource allocation. Slice aware.
Nv_ngp_rat_03	Performance Enhancement of Low-power RRU	Service	Static allocation and limited coordination	Same as above.

8.8 IoT Scenario

8.8.1 Model Architecture/Protocol Stacks

IoT is in the process of maturing from a niche topic to a mainstream topic. There are now many different IoT architectures deployed around the world with varying degrees of industry adoption depending on economics, OPEX, CAPEX, available communications and degree of robustness required. These devices often operate multiple heterogeneous edge links between end sensors and/or actuators and the head-end IoT system. However, Figure 43 illustrates the key IoT communications architecture options which need to be supported in any NGP evolution.

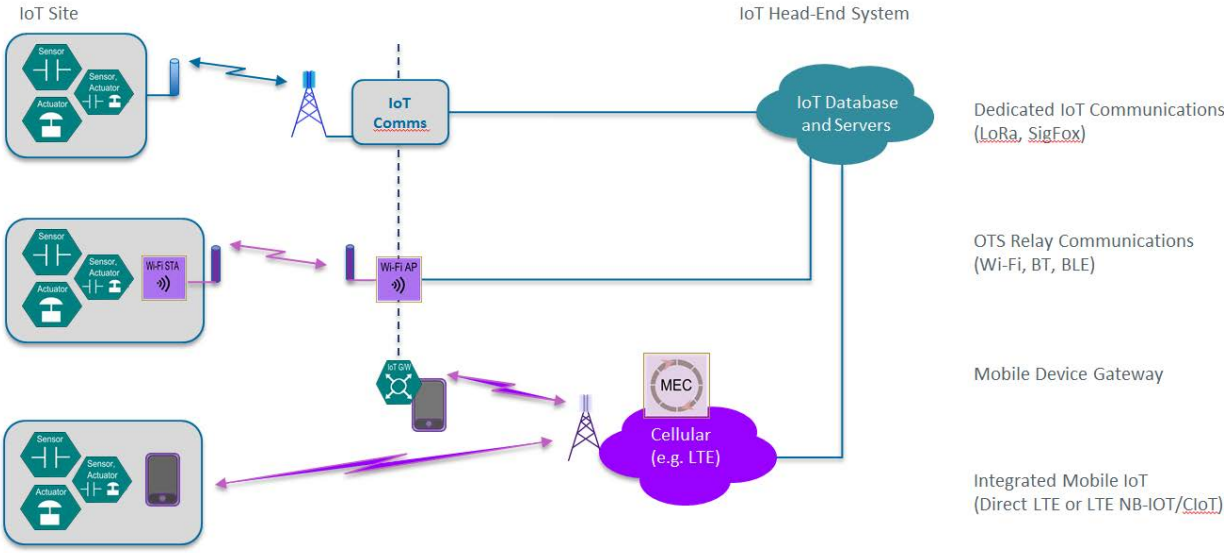


Figure 43: Key IoT Communications Architectures

Figure 43 also illustrates that the use of gateways between different communications links used to provide the composite ETE IoT path is common.

The Figure also illustrates that use of Mobile Edge Computing to deploy either the full or partial IoT application and/or partial database is also an emerging requirement for Next Generation architectures.

The key communications options to be supported are:

i) Dedicated IoT Communications such as LoRa™ and/or SIGFOX™.

ii) Off-The-Shelf Relay Communications such as Wi-Fi™ and Bluetooth™ and/or Bluetooth Low Energy or BLE which is marketed as Bluetooth Smart™.

iii) Mobile Gateway, Cellular Relay using a mobile as the relay Gateway for the IoT information, including the mobile relaying sensor information from itself as well as connected passing IoT equipment. In this case the end IoT communications is whatever is supported on the phone, usually Wi-Fi™, Bluetooth™ and BLE.

iv) Integrated Cellular where cellular device technology is integrated into the IoT sensor actuator system and mat be GSM, GPRS, UMTS, LTE, or a 3GPP NB-IoT/CioT extended range version of the 3GPP standard RAT interface.

8.8.2 Scenario Descriptions

8.8.2.0 Introduction

Most of the IoT scenarios envisaged by NGP ISG for next generation networks are covered by those listed in clause 8.8.4 which references the 3GPP SMARTER use cases. However, some scenario groups from 5GPPP are not covered by SMARTER and these are included in the following clauses.

8.8.2.1 Active Assisted Living (AAL)

AAL refers to the ability for elderly or disabled people to live independently in their own home while being monitored remotely. Various sensors around the house and/or on the person's body or clothing can detect situations such as the person falling down, and alert an operator who can speak to the person and maybe also see them, and summon appropriate assistance if required.

When the system is in operation everything works automatically without any user interface. The remote monitoring application runs in a cloud server and can communicate with, for example, a wristband worn by the person being monitored. A local router, situated for example in a smartphone, may be needed to translate between a local communication method such as Bluetooth Low Energy and the mobile network or Wi-Fi™, but at application level end-to-end security and communication through firewalls are achieved. The communication should be reliable and the power consumption in the wristband low to achieve long battery life. User interaction is only required at system installation. The user, or another trusted person, e.g. a relative, health care personnel or personal assistant, uses a web browser, on another device, to log in to the remote monitoring application and the user has to approve that the application is given access to his/her wristband.

8.8.2.2 Cooperation between factories and remote applications

This use case is illustrated in Figure 44 describes how an IoT application running on a remote computing environment exchanges data with different manufacturing locations through a wide area network in order to optimize operations by monitoring and controlling production lines. In this case, an IoT Gateway located in a factory provides connectivity between the IoT application and controllers (e.g. MES, SCADA, PLC). The IoT Gateway has to adapt different transmission attributes between inside and outside of the factory dynamically.

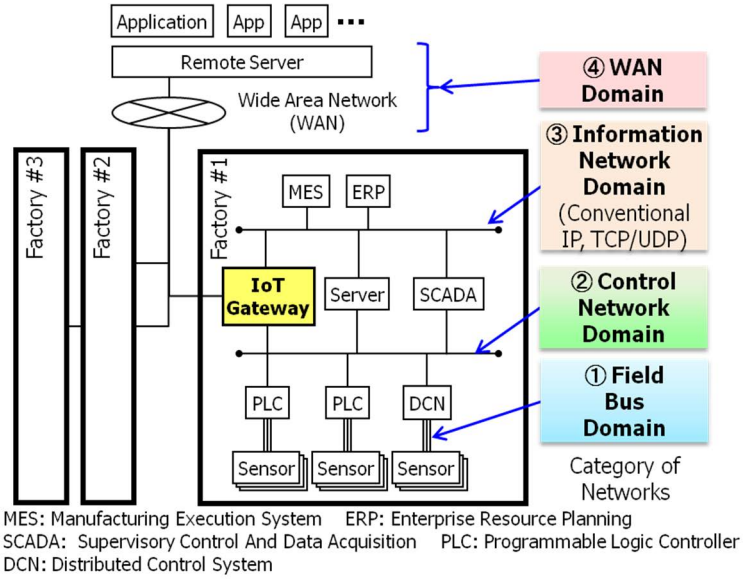


Figure 44: Cooperation between Factories and Remote Applications

8.8.2.3 Smart glasses in industrial applications

Smart glasses enable factory workers to have essential information provided in a "hands free" way so that they can undertake assembly or maintenance operations. With some smart glasses, workers can transmit images from their field of view to remote co-workers who can then give advice. Field and warehouse employees will be among the early adopters for wearable technologies aimed at increasing productivity and safety while reducing employee errors.

8.8.3 Applicable Issues

Issue-01: **Priority and Pre-Emption Support:** There is currently no prioritization or pre-emption support provided in the current internet protocols, for critical IoT groups.

Recommendation-01: NGP should include scalable priority and pre-emption capabilities for IoT services as well as QoS handling.

Recommendation-02: NGPs should include scalable priority and pre-emption capabilities for IoT services as well as QoS handling.

Issue-02: **Scalable Security Support:** Because there are many different types of IoT system that are anticipated to connect to next generation networks with vastly different security requirements, there is a need for a scalable Security Framework.

Recommendation: NGP should provide scalable security features, potentially selectable per IoT device and./or service type as part of a network registration procedure using selectable security profile(s).

Issue-03: **Scalable Addressing Support:** There are likely to be many more IoT devices than people connected to the next generation networks that NGP will have to accommodate, considering massive-IoT roll-out for common monitoring devices to restricted address ranges for more critical IoT communications.

Recommendation: NGP should provide scalable security features, potentially selectable per IoT device and./or service type as part of a network registration procedure using selectable security profile(s).

Issue-04: **Ultra Reliable Low Latency Support:** 3GPP TR 22.891 [i.1] next generation SMARTER use cases and ETSI TR 138 913 [i.25] Next Radio specifications both mention Ultra-Reliable Low Latency Communications (URLLC) support as critical for IoT devices in the next generation of networks that is currently not available in LTE or previous generations of 3GPP cellular networks. There are similar 3GPP requirements for a different class of IoT devices that require Ultra-Reliable, low throughput communications (URC).

Recommendation: NGP should provide support for IoT devices that require ULRCC and URC.

Issue-05: **Interworking Support:** At the lower cost, lower power end of the spectrum of IoT devices, there is a need to interwork different access links together to connect then to the MEC or head-end IoT platform which is not supported in IP today.

Recommendation: NGP should consider IoT support for gateways, bridges and non-NGP interworking.

Issue-06: **Fast Control Loop Support:** Currently the TCP/IP protocols of today do not support open-loop or closed loop fast scanning and/or polling systems such as SCADA based IoT systems.

Recommendation: NGP should consider integrated support options.

Issue-07: There is currently minimal support for the IoT use case group of: **Active Assisted Living.**

Recommendation-01: NGP should ensure that the IoT user can control with whom their information is shared, e.g. only uploading to an authorized health care provider that is approved by the user.

Recommendation-02: NGP should ensure that the IoT device/user is authorized to access a certain health care providers, i.e. protection against fake devices and malicious users.

Recommendation-03: NGP should ensure that IoT confidentiality of information can be assured. No unauthorized entity should be able to get access to the IoT data. This is typically assumed to be solved by operating encrypted transport.

Recommendation-04: NGP should ensure that IoT information integrity is assured. This means that it should not be possible to modify the data being sent.

The system should be reliable in all aspects, for example:

Recommendation-05: NGP should ensure that IoT cloud applications shall be able to detect if any failure occurs, for example if contact is lost with the wristband of a user.

Recommendation-06: NGP should ensure that critical health IoT wearables, for example an IoT care wristband shall be able to self-detect when its battery level is below a certain threshold value and still be able to send an alarm message to the cloud application.

Recommendation-07: NGP should ensure that there are mechanisms in place to be able to support IoT systems when there is a requirement to provide high availability of the communications network, with provision of a suitable back-up means of communication, to avoid unnecessary call-outs.

Issue-08: There is currently minimal support for the IoT use case group of **Co-Operation between Factories.**

Recommendation-01: NGP should provide support for the security of commercially sensitive information.

Recommendation-02: Latency considerations for interworking factory control should be considered by NGP.

Issue-09: There is currently minimal support for the IoT use case group of: **Smart Glasses.**

Recommendation-01: NGP shall provide for the security of sensitive product data including controls for secure access.

Recommendation-02: NGP shall consider the health and safety aspects of Smart Glasses which may be considered as critical systems.

Recommendation-03: NGP shall address latency performance for IoT systems that interact with remote co-workers. The exact latency limits will be application specific.

8.8.4 Applicable Use Cases (from Annex A)

The following IoT-specific use cases are developed from the 3GPP TR 22.891 [i.1] next generation SMARTER use cases.

SMARTER referenced use cases 1a, 1b, 20, 21, 24, 25, 40, 42, 43, 44, 45, 46, and 59 are identified as having notable IoT scenario dependence and/or are required in order to support next generation IoT enhancements.

8.9 Energy Efficiency

It is already recognized in the Digital Single Market project of the European Commission through projects such as ECONET, that it is imperative that the most energy efficient transport network possible be designed for next generation networks by minimizing protocol overhead, for example, by minimizing multiple layers of extensive headers in the same protocol stack.

There are several potential areas where energy can be wasted in network transmission stacks, as follows:

Issue-01- **Processing Impact:** Each separate address in multiple layers requires more complex IO devices such as ASICs which all consume power. The more layers, the more addressing processing and the more power consumed.

Compression is also a power hungry processing overhead in terms of header compression and payload compression.

Recommendation: NGP shall minimize the need for complex address, header, compression and tunnelling processing in handling network protocols.

Issue-02: **Header Storage:** Handling network protocol headers, requires that portions of each packet be held in memory or buffer structures; the more levels of information which need to be held, the more memory space will be required, which is directly related to the cost of operation and cost of manufacture/provision of such memory.

Recommendation: NGP shall minimize the need for header storage.

Issue-03: **Protocol Efficiency:** The ratio of useful data in the payload to overhead has a direct financial impact on communication links; these links are of finite capacity and hence have a finite cost-per-unit-data that can be calculated.

The capacity used to transport information as compared to the overhead which is unavailable for use by a customer, but required to transmit is often expresses as a good-put efficiency and can be related to cost to transmit payload data. This is a major cost driver and therefore inefficient overhead degrades the overall system efficiency.

Recommendation: NGP protocols shall minimize header complexity and overhead and build in header scalability according to context and protocol usage type.

Whilst there are tangible costs associated with providing power to components in the transport network, there are various less-tangible costs associated with the provision of such capacity.

Issue-04: **Manufacturing Cost:** There are manufacturing costs in-terms of raw materials cost, factory power and environmental impact for the extraction of the electronic components used in communications systems.

Designing and deploying systems is more expensive where complex systems with multiple levels of interaction are concerned.

There will be an associated operational cost over the life-time of the system due to the complexity of the solution.

Recommendation: NGP should minimize protocol complexity and multi-level manufacturing required to implement protocols.

8.10 eCommerce

Whilst eCommerce is undoubtedly one of the main reasons that users access the internet today it is not envisaged that the ETSI, NGP, ISG will produce any specific scenarios that highlight issues with the support of eCommerce for Next Generation Protocol architectures or that the ETSI, NGP, ISG will make any recommendations for eCommerce support.

However, a general improvement in the security capabilities on offer to a user when operating the network whilst performing eCommerce are considered to be addressed in the Security clause 8.2 of the present document.

As to security support specifically for eCommerce applications operated over the NGP architecture, it is widely thought that they should not rely on the services of the Network to protect them.

8.11 Mobile Edge Computing (MEC)

8.11.0 Introduction

Mobile Edge Computing (MEC) is a new technology which is currently being standardized in the ETSI Industry Specification Group (ISG) MEC.

Mobile Edge Computing provides an IT service environment and cloud-computing capabilities at the edge of the mobile network, in close proximity to mobile subscribers, see Figure 45, in order to reduce latency, ensure highly efficient network operation and service delivery, and offer an improved user experience.

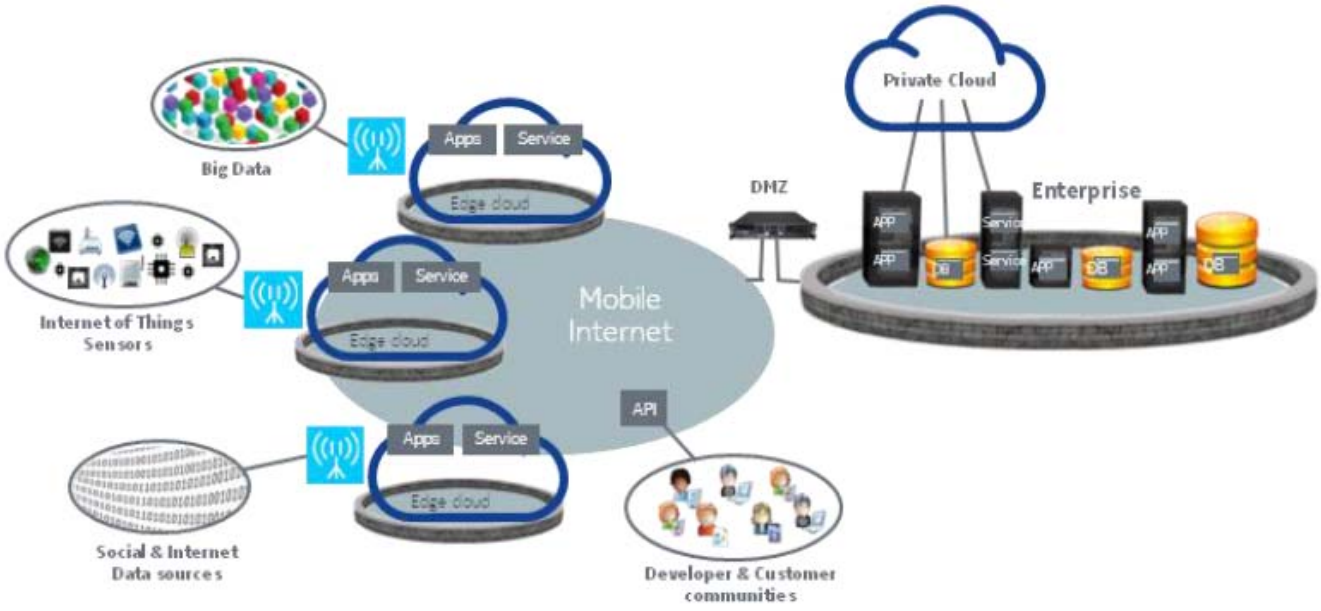


Figure 45: Improved QoE with Mobile Edge Computing in close proximity to end users, from reference [25]

8.11.1 Model Architecture

The ETSI Industry Specification Group (ISG) on Mobile Edge Computing (MEC) produces normative Group Specifications that will enable the hosting of applications, including the third party applications, in a multi-vendor MEC environment. This ISG was launched in December 2014, and plans to deliver the first set of specifications within 2 years. The initial scope of the ISG MEC focuses on use cases; it specifies the requirements and the reference architecture, including the components and functional elements and the reference points between them. The Group Specifications covering the requirements and the reference architecture, ETSI GS MEC 002 [i.21] and ETSI GS MEC 003 [20], correspondingly, were published in March 2016. The ISG has published a set of GSs covering the management of the system, host and the mobile edge applications, as well as the application programming interfaces for offering and consuming services in the mobile edge host:

- ETSI GS MEC-IEG 005 [i.22] - Proof of Concept Framework, specifying the process and criteria that a Proof of Concept demonstration should adhere to.
- ETSI GS MEC-IEG 004 [21] - Service Scenarios, which presents a number of examples of service scenarios, business and consumer benefits which can be enabled by Mobile Edge Computing.

Figure 46 and Figure 47 show the Mobile Edge Computing Framework and Mobile Edge Computing Reference Architecture, respectively [20].

In particular, Figure 46 shows the framework for Mobile Edge Computing consisting of the following entities:

- Mobile Edge Host, including the following:
  - mobile edge platform;
  - mobile edge applications;
  - virtualisation infrastructure;
- Mobile Edge System Level management;
- Mobile Edge Host level management;
- External related entities, i.e. network level entities.

Figure 47 shows three groups of reference points that are defined between the system entities:

- Reference points regarding the mobile edge platform functionality (Mp);
- Management reference points (Mm);
- Reference points connecting to external entities (Mx).

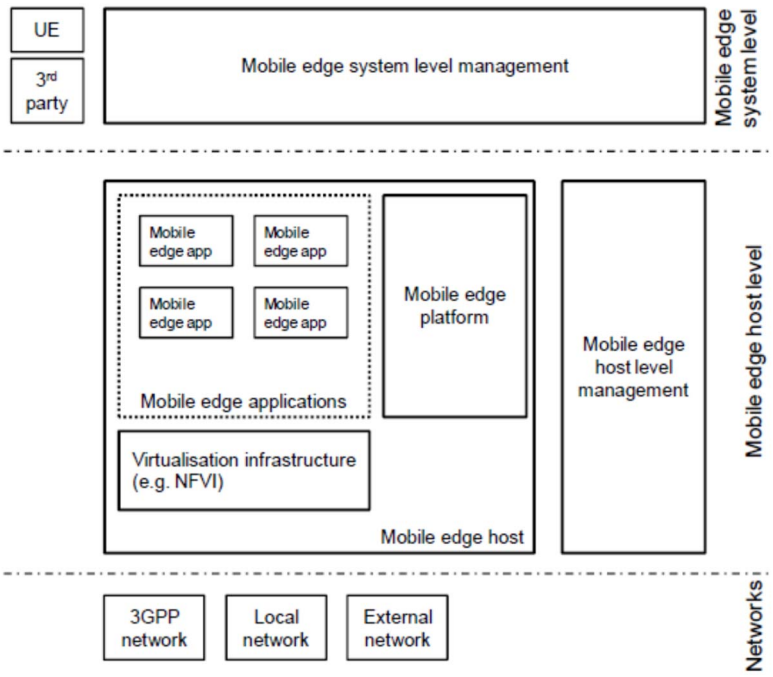


Figure 46: Mobile Edge Computing Framework, from ETSI GS MEC 003 [20]

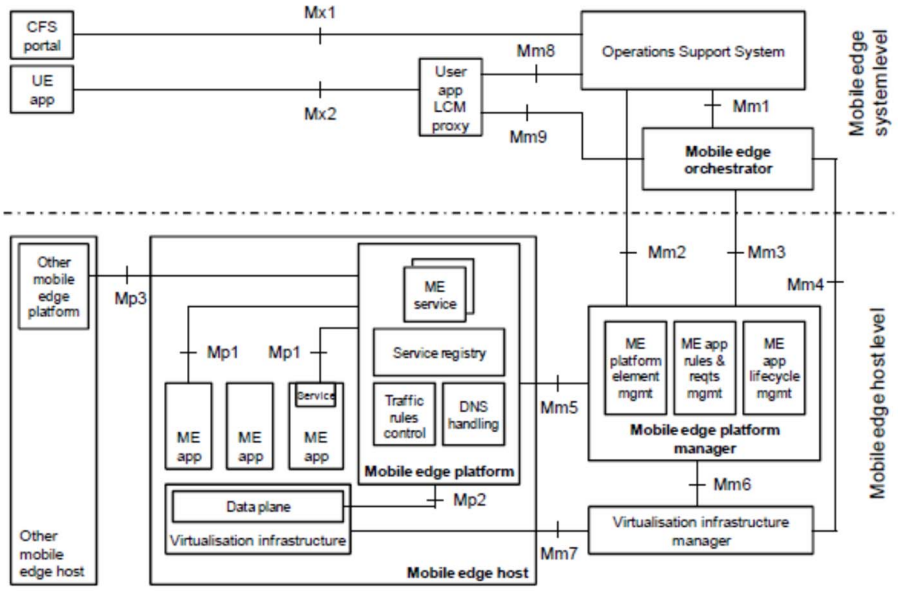


Figure 47: Mobile Edge Computing Reference Architecture, from ETSI GS MEC 003 [20]

8.11.2 Applicable Issues and Recommendations

MEC can be seen as an emerging technology that can be used to support any Next Generation Protocol development by enabling processing, storage and cloud computing capabilities at the edge of the network.

Issue-01: There is currently no concept or the ability for either the user to select or the network to direct services to be served either at the edge of a system or in the core of a system. With the recent notable increase of non-wired access to the internet, providing service at the edge is becoming an essential capability for efficient QoE provision to users.

Without this capability it is difficult to provide good end to end network resource provision and life cycle management for such services as UHD video and VR/AR services.

Recommendation: NGP should provide protocol support that enables edge cloud platform capabilities that can offer such capabilities as caching, pre-fetching and other edge hosted capabilities for such services as video, application and VR optimization.

Issue-02: Currently the edge nodes themselves do not possess any form of standard Edge computing capabilities which makes early Edge Computing capabilities proprietary and difficult to integrate with the internet.

Recommendation: NGP should provide access agnostic capabilities that enhance next generation wireline and wireless edge nodes for supporting the Edge computing, storage and optimization. Such features to be considered in NGP are:

1) edge-cloud capabilities; and

2) differentiating conventional traffic from traffic related to cloud applications, e.g. computation offloading and storage.

8.11.3 Applicable Use Cases

8.11.3.0 Introduction

MEC is mentioned several times in the SMARTER document as part of a number of Use Cases. For clarity, in the present document, specific MEC use cases are introduced in the following clauses.

8.11.3.1 Case 1: Video Stream Analysis service

This scenario refers to video stream analysis and is based on ETSI GS MEC-IEG 004 [21]. Video-based monitoring currently requires either sending video streams to a server or for video processing to be done at the same site as the camera. Both methods are costly and inefficient when compared to processing video data at a MEC server in order to extract meaningful data from video streams, see Figure 48. Subsequently, the valuable data can then be transmitted to the application server without the need to transport high data rate video streams.

The key benefit of this scenario is that performing the analysis locally, i.e. close to the edge of the network, mitigates the need to transmit high-data rate video streams when only small pieces of information are required to be extracted from these video streams.

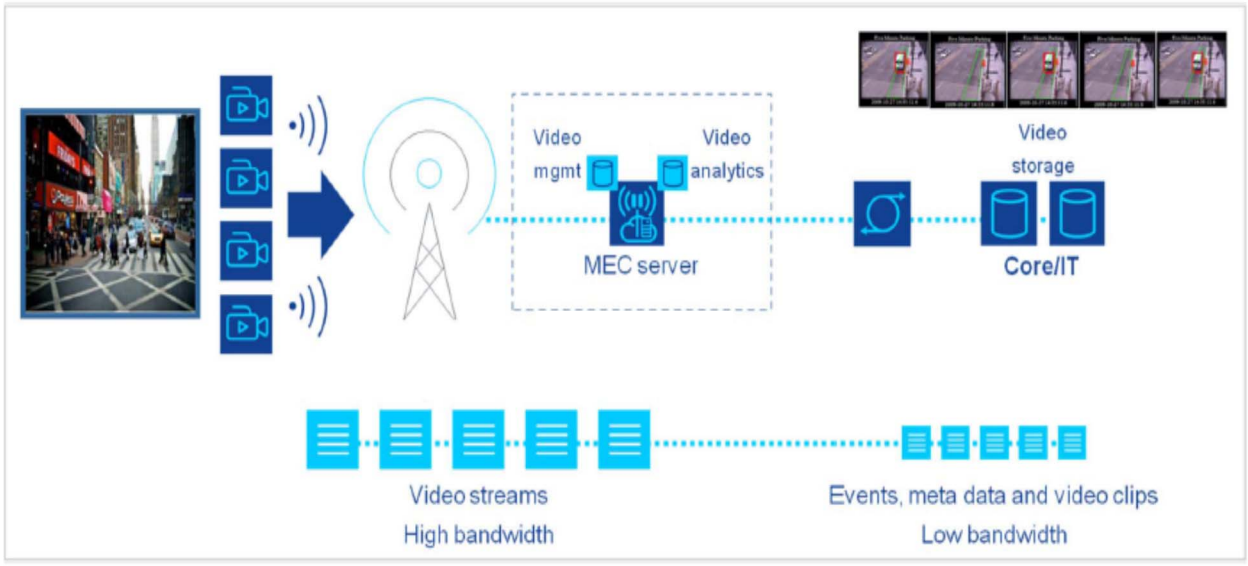


Figure 48: Video Stream Analysis, based on ETSI GS MEC-IEG 004 [21]

8.11.3.2 Case 2: Augmented and Virtual Reality service

This scenario refers to augmented and virtual reality services and is based on ETSI GS MEC-IEG 004 [21]. Typically, an augmented reality service supplements a user's experience by providing additional information to the user about what they are currently experiencing. For example, augmented reality can enhance the experience of a visitor to a museum or another point of interest. This is currently done, by requiring an application to analyse the output from a device's camera and/or a precise location in order to supplement a user's experience when visiting a point of interest. After enhancement such information, the application can provide additional information in real-time to the user.

The key benefit of this scenario is that augmented information pertaining to a point of interest is highly localized and thus hosting the information locally is advantageous compared with hosting in the cloud.

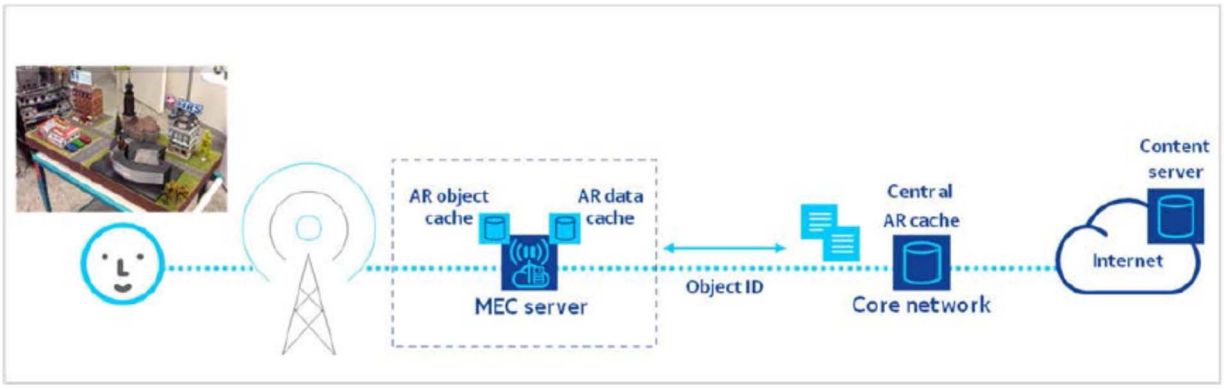


Figure 49: Augmented Reality, from ETSI GS MEC-IEG 004 [21]

8.11.3.3 Case 3: Assistance for intensive computation service

This scenario refers to Assistance for intensive computation service scenario and is based on ETSI GS MEC-IEG 004 [21]. Currently, several wireless devices or sensors are made to be as low cost as possible but are required to remain operational for a long period of time.

Such a wireless device may also require further instruction or feedback based upon the information it feeds to the application or service. As a result of the low cost, computational power is often sacrificed. By performing such computation off-board, i.e. at a MEC server located in the neighborhood, the wireless device or sensor can increase the battery life of remote devices.

The key benefit of this scenario is that the offload computationally intensive data processing to a MEC server, can improve the performance of a device with low processing power and also improve battery performance.

8.11.3.4 Case 4: IoT Gateway service

This scenario refers to IoT Gateway service scenario and is based on ETSI GS MEC-IEG 004 [21]. Currently, most of the IoT devices and sensors are constrained devices from the point of processor and memory. Because of the nature of such IoT devices and sensors being connected, a real time capability is required and a grouping of the IoT devices and sensors is needed for efficient service. It is beneficial to use an IoT gateway to aggregate various IoT device messages connected through the mobile network close to the devices, see Figure 50. This will provide an analytics processing capability and a low latency response time.

The key benefit of this scenario is that it supports a low latency aggregation point to manage the various protocols, distribution of messages and for the processing of analytics required by IoT devices and sensors. The MEC Server provides the capability to resolve these challenges.

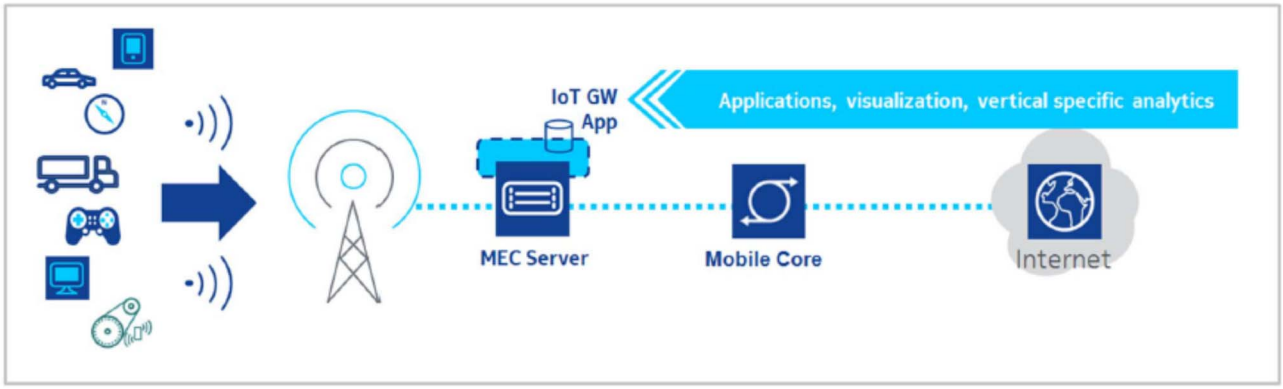


Figure 50: IoT Gateway, from ETSI GS MEC-IEG 004 [21]

8.11.3.5 Use Case 5: Connected Vehicles service scenario

This scenario refers to Connected Vehicles service scenario and is based on ETSI GS MEC-IEG 004 [21]. Currently, the number of cars and other vehicles becoming more 'connected' using technologies such as Dedicated Short Range Communications (DSRC) for short distance and Long Term Evolution (LTE) for long distance connectivity is increasing. The support of communication of vehicles and road side-sensors is intended to increase the safety, efficiency, and convenience of the transportation system, by the exchange of critical safety and operational data. However, as the number of Connected Vehicles increases and the situations where the use evolves, the volume of data will continue to increase along with the latency requirements. By storing and processing the data centrally may satisfy the requirements of some use cases, but it can be unreliable and slow for all uses. A MEC server located at Road Side Units and/or close to them at the edge of the mobile network can be used to store and process such data, see Figure 51.

The key benefits of this scenario are that by locating the MEC server at Road Side Units and/or close to them at the edge of the network can enhance the performance and reliability of services for Connected Vehicles.

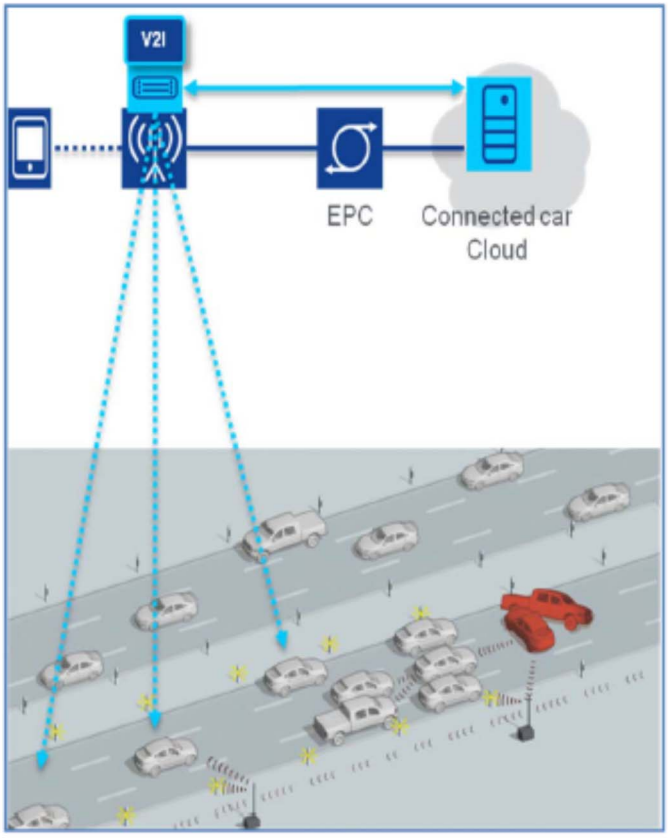


Figure 51: Connected Vehicles, from ETSI GS MEC-IEG 004 [21]

8.11.4 Scenario Targets

Table 9 details the KPIs for improvement of this Scenario as a result of the development of NGP new addressing system.

Table 9: KPI's for Scenario - 11

KPI Name	Description	Measured feature	Current behaviour	Target Value/ behaviour
NGP_MEC_01	Intelligent video acceleration service scenario		ETSI GS MEC-IEG 004 [21]	
NGP_MEC_02	Video Stream Analysis service scenario		ETSI GS MEC-IEG 004 [21]	
NGP_MEC_03	Augmented and Virtual Reality service scenario		ETSI GS MEC-IEG 004 [21]	
NGP_MEC_04	Assistance for intensive computation service scenario		ETSI GS MEC-IEG 004 [21]	
NGP_MEC_05	IoT Gateway service scenario		ETSI GS MEC-IEG 004 [21]	
NGP_MEC_06	Connected Vehicles service scenario		ETSI GS MEC-IEG 004 [21]	

8.12 Mission Critical Services: PSC and PUC

8.12.0 Introduction

For several generations of 3GPP releases, features for the support of Mission Critical Services have been defined to support both Public Safety Communications (PSC) and Public Utility Communications.

However, due to commercial reasons, mainly of cost of implementation of mobiles and infrastructure supporting these services little has been adopted over commercial mobile services.

With the advent of virtualisation and network slicing as introduced in some of the early 5G technical reports such as 3GPP TR 22.891 [i.1] and 3GPP TR 23.799 [i.2] it is likely that a truly diversified set of services is able to be realized over one cellular network infrastructure, but with substantially different services and QoS, QoE per slice of the network.

As such, the present document notes that the scenarios already defined in previous releases of 3GPP for Mission Critical: PSC and PUC, should be introduced for consideration when defining Next Generation Protocols. See ETSI TS 122 280 [29] and 3GPP TR 22.862 [i.34].

8.13 Drones, Autonomous and Connected Vehicles

8.13.0 Introduction

This clause describes a set of next generation, network connected vehicles; lists use cases applicable to next generation protocols; and assesses whether such vehicles present any unique requirements.

8.13.1 Model

Several studies have shown that driving automation can be divided in several levels. The Society of Automotive Engineers have identified six levels of automation in standard J3016, from "no automation" to "full automation", see [30]. These automotive levels are broadly consistent with the National Highway Traffic Safety Administration's "Preliminary statement of policy concerning automated vehicles" (2013), see reference [i.35], and are referenced by the "5G Automotive Vision" 5GPP white paper [i.36].

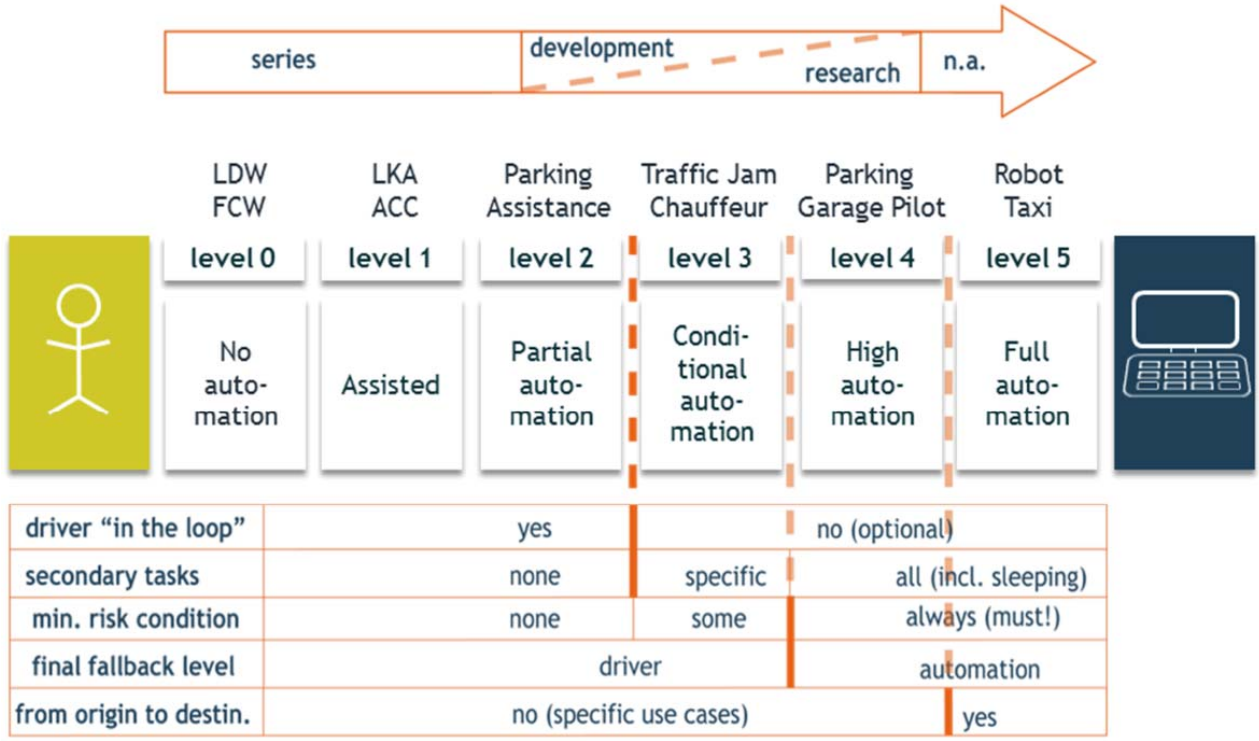


Figure 52: Levels of Vehicle Automation (source reference: [i.38])

For the purposes of the present document, levels 4 and 5 will be considered primarily, in which a human operator is not expected to be involved in piloting decisions or operations.

8.13.2 Scenarios

The scenarios for Drones, Autonomous and Connected Vehicles, for 5G, are specified in 3GPP TR 22.886 [i.37].

3GPP TR 22.886 [i.37] includes service requirements for Vehicle to X (V2X) scenarios of use, from which the following use cases for Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) scenarios are derived:

- Automotive: sensor and state map sharing
- Collective perception of environment
- Cooperative collision avoidance (CoCA) of connected automated vehicles
- Information sharing for limited automated driving
- Information sharing for full automated driving
- Information sharing for limited automated platooning
- Information sharing for full automated platooning
- Video data sharing for assisted and improved automated driving (VaD)
- Changing driving-mode
- Emergency trajectory alignment
- Intersection safety information provisioning for urban driving
- Cooperative lane change (CLC) of automated vehicles
- Considerations on network slicing

8.13.3 Applicable Issues and Recommendations

Issue-01: Autonomous vehicles will be expected to be self-piloting, including in situations with network communications failure. Therefore, the **network coverage, reliability, latency and throughput** requirements for data transiting to and from the vehicle will not be based on piloting data, but rather other data scenarios: for example, video feeds as a drone inspects an area may be critical to be sent in high-resolution without stalling.

Recommendation: Across the range of Next Generation Vehicle applications, there may be requirements for high network capacity/throughput and low latency; however, these are not considered as core use cases. Latency is not a critical factor for the piloting of an autonomous drone with a pre-programmed flightpath and self-piloting capabilities; but if the drone is transmitting (for example) video or real-time sensor readings then it may become so. Similarly, high throughput may be required if the drone is carrying a camera to transmit live video. Therefore, the network coverage, reliability, latency and throughput requirements for Autonomous Vehicles should take into account the non-piloting communication scenarios between the Vehicle and the network.

Issue-02: Supplementary to Issue-01, in the edge case where an Autonomous vehicle is being **piloted remotely via a network connection**, then any delay or break in communication can result in a safety risk.

Recommendation: A remote-piloted Autonomous Vehicle should be migrated to a reliable, low latency network slice; or the network should be able to announce if such a slice is not available for the current vehicle location.

Issue-03: **Vertical coverage:** Cellular networks are typically tuned to focus their radio coverage at around the height an average human will hold their phone during a conversation: between five or six feet above ground level. Road-based autonomous vehicles may receive cellular signals at a lower height (maybe three to four feet above ground), but aerial drones will operate up to approximately 500 metres above ground, with a regulatory limit imposed by a given territory's airspace policies.

Recommendation: Network operators will need to consider how to achieve coverage at increased altitudes: both in terms of physical deployment (mast height, mast location and antennae mounting position) and in terms of spectrum licence terms - namely, at what altitude limit they can transmit and receive radio. Aquatic vehicles may operate at lower (sub-street level) altitudes and hence may also require repositioning of network radio antennae for best performance and coverage.

Issue-04: **Failsafe:** Autonomous Drones present an important and unique consideration towards next generation protocols: physical damage, to themselves, people, or their environment, may be incurred as a result of communications failure. This is because these vehicles gain momentum from thrust and/or the potential energy resulting from altitude gain. Crucially they are unmanned, meaning no direct human intervention in case of loss of network-sourced instructions' hence the risk of uncontrolled inertia of the moving or hovering vehicle.

The vehicle should have a failsafe mechanism, determining how the vehicle should act in the event of a communications failure, or failure of on-board diagnostics (such as an inability to read the remaining power available to the vehicle). 'Heartbeat' mechanisms are available in autonomous drone application protocols today, such as Mavlink (http://qgroundcontrol.com/). This allows the drone to infer connection loss and implement a recovery process, for example:

NOTE 1: Mavlink is an example of a suitable product available commercially. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of this product.

1) stop and loiters for a few seconds (to see if connectivity is restored);

2) climb to a minimum safe 'return to base' altitude (if below that altitude);

3) return to launch coordinates (or another pre-configured coordinate);

4) wait a few seconds over the home location and then lands automatically.

External applications may be available to the remote pilot to avoid connection blackspots, physically congested flightpaths, and local environmental conditions; thus reducing the need for a failsafe to be applied:

Recommendation: Although failsafe and flightpath planning are applications, there is a requirement on the network layers to provide guaranteed connectivity, i.e. extremely high reliability, to allow remote pilot awareness of the drone's status. For network protocols, the contribution to failsafe may be in helping prevent the failsafe scenario from occurring by informing flight planning applications of any network blackspots or congestion.

Issue-05 **Machine ethics:** In a situation where an autonomous vehicle will collide with either one or another pedestrian, for example a pensioner and a child, how will the vehicle choose which to collide into and which to avoid? Driverless Future [i.39] argues that this is an irrelevant question, because such situations are rare; and that there is no definitive solution to the dilemma for even human drivers to apply. It also argues that the general approach towards ethics for autonomous vehicles should be to avoid what is ethically wrong (which is easier to define) rather than to choose what is ethically right.

Recommendation: Network operators should follow any emerging regulations around machine ethics, however it seems that should any such regulation be enforced, that it would be the autonomous vehicle making the decision in situ (due to the low-latency requirement) without network assistance.

Issue-06 **Geofencing**

Figure 53: Example of drone exclusion notice. Note the out-of-band permission process

Geofencing here refers to policies around use of next generation vehicles within a geographical boundary. Typically, this would apply to airborne (or possibly aquatic) vehicles, although it may apply to off-road capable land vehicles.

Certain areas may prohibit or restrict autonomous drones in their airspace - that is, any airspace over that land not regulated by that territory's aviation authority. This includes privacy, but also safety where the area has physical hazards (wind turbines, power lines, forests, etc.).

Since cellular radio coverage is not strictly tied to local boundaries - such as a farm, estate or home - it is possible to use cellular networks to pilot a drone into such areas. Network enrolment alone is therefore not sufficient to admit or deny drones (although see note on roaming below).

Recommendation: Network operators should broadcast geophysical boundaries (including altitude) from radio access points, and applications on the Autonomous Vehicle should process these for access restrictions. A network operator may also offer to broker access to an area, including via:

1) Announcement: the drone makes itself known to the radio access point (analogous to the 'listen out squawks' available to UK pilots and Air Traffic Controllers) which triggers an access request dialogue.

2) Policy broadcast: the landowner makes the policy available for drones to discover.

3) Conditions for entry: for example, payment-based, or restrictions, such as no filming.

Cellular roaming across legislative boundaries (https://www.inverse.com/article/21882-drone-data-plans) can bring the drone into areas with different general airspace regulations. In this case it may be possible for Network Enrolment to include and enforce geopolicing, as the roaming handover is set up.

NOTE 2: INVERSE is a trademark of Full Stack Media and is an example of a suitable product available commercially. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of this product.

Issue-07: **Privacy:** This does not refer to privacy of the end user - the remote pilot - but rather the privacy of citizens in the area traversed by a drone carrying a monitoring device (camera, microphone, Wi-Fi sniffer etc.).

Recommendation: Privacy policy may be realized by the 'Geofencing' section above; and may extend that to deter a 'stalking' scenario in which a drone follows an individual.

Issue-08: **Security:** Piloting instructions for Next Generation Vehicles will require a stronger security than other data (e.g. entertainment streaming) due to the physical harm risked due to data hijacking or corruption. In addition, Next Generation Vehicles transporting human passengers require higher safety guarantees than unmanned vehicles.

Recommendation: If the network is transporting piloting instructions, the Next Generation Vehicle should be able to request higher reliability, lower latency and stronger security for those piloting instructions than for any non-critical data (such as infotainment). An Autonomous Vehicle transporting human passengers in this scenario should be granted the highest level of security and reliability.

Issue-09: **Resource contention:** Where Connected vehicle entertainment services are delivered over the same access network as vehicle/piloting data, then resource contention may occur.

Recommendation: A network policy may choose which data to deliver and which to delay. The same may apply within vehicle/piloting data; for example, delivering an engine status update to the car manufacturer for Big Data analysis may be seen as less critical than delivering data regarding local road conditions.

Issue-10: **Auditing:** The actions of a drone or autonomous vehicle that has been involved in an accident may be crucial in law enforcement or insurance investigations. This includes a distinction between actions directly performed by the human operator (where relevant), the actions based on on-board sensor input processing, and instructions or communication failures from the network; to inform liability decisions.

Recommendation: A network should be able to isolate the flows pertinent to a particular vehicle journey for liability purposes.

8.13.4 Applicable Use Cases

8.13.4.0 Introduction

This clause expands on some of the scenarios related to 3GPP TR 22.886 [i.37].

Due to the inherent mobile nature of vehicular cases, wide **coverage** and **mobility** across access points (with no or low latency or data loss) are common to all the categories. **Reliability** and **security** is very important for piloting data sent over the network, since physical harm can occur due to connection failure, data corruption or connection hijacking.

8.13.4.1 Hazardous operations

Autonomous Drones can operate where it is hazardous for a human operator. Cases include medium to high-altitude inspections of building sites or infrastructure requiring maintenance; undersea inspection or surveying of the seabed for cabling or oil rig supports; or police/military drones that can search for and defuse explosives.

Drones have been used to extend cellular coverage in areas where it would not be practical, or cost-effective, to perform through more traditional means; such as extending LTE coverage at rural events: http://about.att.com/innovationblog/drones_new_heights. Other investigations mooted include the delivery of power wirelessly to difficult-to-access sensors on bridges: https://eandt.theiet.org/content/articles/2016/10/wirelessly-powered-drone-enables-indefinite-flight-time/.

NOTE: AT&T is a trademark for the American Telephone & Telegraph company in the US, 'The IET' is a registered trademark in the UK for the Institute of Engineering and Technology, Senseable is a trademark of AMS, in Amsterdam these are examples of suitable products available commercially. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of these products.

8.13.4.2 Driverless vehicles

https://www.asirobots.com/autonomous-solutions-inc-and-cnh-industrial-unveil-concept-autonomous-tractor/ describes a tractor that can operate as manned or unmanned, with remote network control from the farmer: autonomous boats are being investigated for logistics, human transport, water measurements and even physical infrastructure (temporary bridges formed by connected boats) http://senseable.mit.edu/roboat/. Connecting these autonomous vehicles in large rural areas and urban waterways may require coverage improvements from network operators.

Driverless cars have been trialled in multiple locations worldwide. Data may be shared:

- vehicle-to-vehicle, as broadcast. Information shared may include intended manoeuvres or warnings, such as upcoming road potholes or upcoming congestion;
- vehicle-to-edge-network;
- vehicle-to-macro-network, for less latency-sensitive information.

Autonomous vehicles may interact with other vehicles, and static transport network access points (e.g. roadside radio cells) via broadcast. Since the information exchanged involves indication of piloting actions (turning, braking, etc.) then there will be a low latency requirement to ensure it can be received, processed and acted on in time.

NOTE: ASI is a trademark of Autonomous Solutions Inc., Senseable is a trademark of AMS, in Amsterdam, MIT is a trademark of Massachusetts Institute of Technology, these are examples of suitable products available commercially. This information is given for the convenience of users of the present document and does not constitute an endorsement by ETSI of these products.

8.13.4.3 Automated Convoys ('platooning')

Automated vehicle convoys involve an automated master vehicle leading multiple slave vehicles, for example a convoy of trucks. This scenario is referred to as 'platooning' in 3GPP TR 22.886 [i.37]. The master vehicle is responsible for communicating piloting instructions to the slave group, hence the slaves are not autonomous. The benefits include energy savings from streamlining, as the convoy can be strictly controlled to reduce the gap between vehicles. The master vehicle will need a long-lived, low-latency secure communication session with the slaves to persist across the journey, and a fallback mechanism in case of connection failure with convoy vehicles.

8.13.4.4 Connected vehicles

3GPP TR 22.886 [i.37], clause 8.11.3.5 describes vehicles sharing driving condition data with roadside servers, to improve the transport network (note this can also apply to aquatic vehicles interacting with buoys or shore servers).

Passengers of connected vehicles may enjoy entertainment services on in-vehicle clients (screens, visors, headphones), which could include media delivered over a macro network, or local information services (sights, information about shops etc.) delivered from an edge network. These may require high-capacity, high-throughput, short-range access networks with rapid mobility handover.

Rapid handover between access nodes is defined for example at cellular edge nodes mounted on street furniture as a car travels along a road that should persist sessions with minimal delay.

8.14 URLLC: Ultra-Reliable and Low Latency Communications

8.14.0 Introduction

In the 5G era, many services (e.g. V2X, Local UAV Collaboration, and Industrial Factory Automation) require Ultra-Reliable and Low Latency Communications (URLLC) which is a term introduced to 3GPP in specification 3GPP TR 38.801 [i.28].

According to 3GPP TR 38.801 [i.28] URLLC is about providing the high likelihood of delivering error free packets through the 3GPP system within a delay bound per packet/message. Furthermore, ETSI TR 138 913 [i.25] defines the target delay bound in terms of user plane latency of less than 0,5 ms for UL, and 0,5 ms for DL for successful delivery of an application layer packet/message between new radio protocol layer 2/3 SDU ingress a degress points.

Low Latency is the time cost by packets transmission between UE and application node within a low delay bound. Table 10 shows requirements for typical URLLC services.

The use cases for this scenario section are referenced from 3GPP TR 22.891 [i.1] and the scenario target values quoted are referenced from 3GPP ETSI TS 122 261 [28].

The parameters quoted here are taken from the annexe of 3GPP ETSI TS 122 261 [28].

Table 10: Latency and reliability requirements for typical URLLC services - 3GPP TR 38.801 [i.28] and ETSI TR 138 913 [i.25]

USE Case Type	Reliability and Latency Requirement	Mobility
Industrial Factory Automation	a residual error rate of < 10 - 9; latency < 1 - 10 ms	Low
V2X	a residual error rate of < 10 - 9; latency < 15 ms	High
Local UAV Collaboration	Position accuracy within [10 cm], Latency < 10 ms	High

The current mobile network protocol cannot meet the requirements of URLLC, and much effort can be done to achieve this. In addition to optimization of encoding, error correction, retransmission mechanism and more in the layer 1 or 2 within radio network, issues also need to be concerned by layer 3 and above. This paper mainly focuses on the following challenges faced by layer 3 or above:

1) How to reduce interruption latency caused by handover. Mobile will trigger the process of handover. In current 3GPP mobile network architecture, about 50 ms interruption comes with handover, which will increase the end-to-end latency.

2) How to reduce interruption latency caused by mobility of application. In order to keep the application node closed to UE, mobility of application should be supported. The migration may cause the application node out of service for a short period, which will increase the end-to-end latency.

3) How to ensure deterministic processing delay in network nodes (such as virtual switches). DiffServ is a set of end-to-end quality of service (QoS) capabilities, which can only ensure the relative delay and is not deterministic. Processing delay is a component of E2E latency, so indeterminate former may enlarge the later.

4) How to sustain high reliability within low latency. Wireless packet loss is unavoidable due to unstable wireless channel. Retransmission can ensure reliability at the expanse of enlarging latency. So it is necessary to consider how to balance packet loss and latency.

NGP should take into account the characteristics of mobile network so as to meet the requirements of URLLC.

8.14.1 Model architecture

This scenario is described so as to identify the issues that should be considered for current and evolving 3GPP mobile network architecture providing URLLC for UE with respect to next generation protocols. The two architectures describe as following.

Figure 54 shows the current Rel-13 mobile network architecture. Detailed information about network elements and reference points for this architecture are detailed in ETSI TS 123 401 [27].

Figure 55 shows a summary of the evolving 3GPP, 5G architecture as discussed in 3GPP TR 23.799 [i.2]. The 5G architecture is now being designed in specification ETSI TS 123 501 [32]. Figure 55 illustrates the decoupling of control and user plane envisaged for 3GPP, 5G which will bring more architectural flexibility.

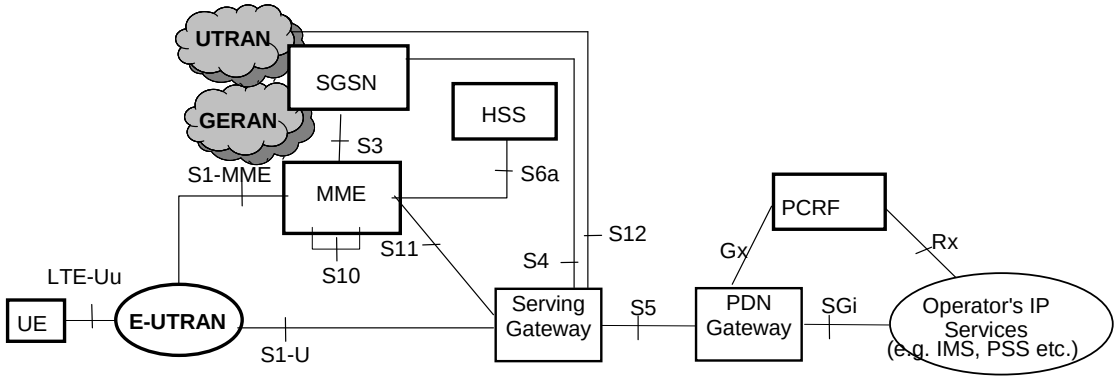
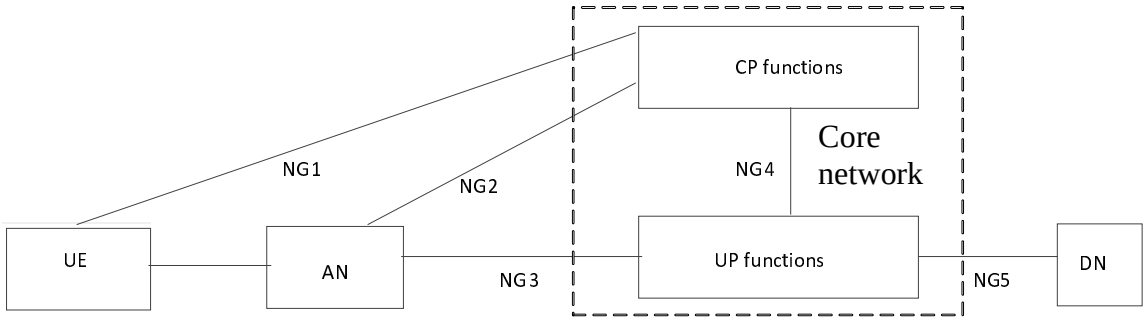


Figure 54: Architecture in 3GPP Rel-13, ETSI TS 123 401 [27]



NOTE: Where:

UE:	User Equipment
AN:	Access Node
DN:	Data Network
CP function:	Control plane function
UP function:	User plane function
NG1:	Reference point between the UE and the CP functions.
NG2:	Reference point between the AN and CP functions.
NG3:	Reference point between the AN and UP functions.
NG4:	Reference point between the CP functions and UP functions.
NG5:	Reference point between the UP functions and a Data Network (DN).

Figure 55: Architecture Proposed for 3GPP '5G'/TR 23.799 [i.2]

8.14.2 Scenario description

8.14.2.0 Scenarios Introduction

There are four scenarios in this clause. The first three are about low latency and the last one is about high reliability.

8.14.2.1 Handover interruption caused by UE's mobility

The user's movement will lead to handover processes triggering and therefore this will affect the end-to-end latency. For example, in the handover procedure in the existing 3GPP Rel-13, the UE first disconnects from the source AN and then connects to the target AN, depicted in Figure 56. During the handover procedure, the link will be break, and the end-to-end latency increases by about 50 ms interruption latency [i.29]. This is intolerable for URLLC. Minimizing the handover interruption latency is one of the challenges for the existing mobile network in serving URLLC.

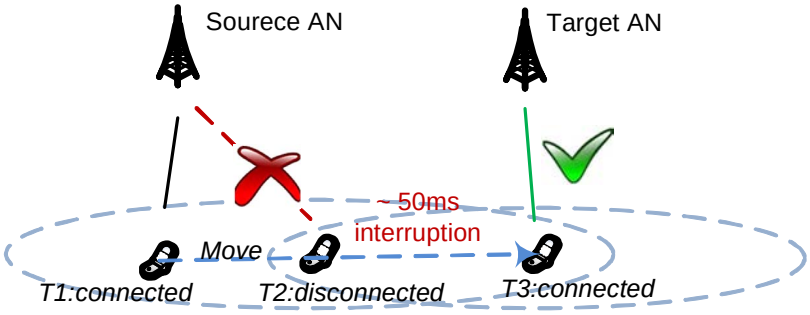


Figure 56: Handover interruption: *about 50ms between T2 and T3*

Table 11: Minimum /Typical radio access latency components during handover [i.28]

Component/ Step	Description	Time (ms)
7	RRC Connection Reconfiguration Including *mobilityControlInfo*	15
8	SN Status Transfer	0
9.1	Target cell search	0
9.2	UE processing time for RF/baseband re-tuning, security update	20
9.3	Delay to acquire first available PRACH in target eNB	0,5/2,5
9.4	PRACH preamble transmission	1
10	UL Allocation + TA for UE	3/5
11	UE sends RRC Connection Reconfiguration Complete	6
	Minimum/Typical Total delay [ms]	45,5/49,5

8.14.2.2 Interruption Latency caused by mobility of application

Traditionally, the application node deploys in the internet. So, in the framework of 3GPP Rel-15 network architecture, the traffic between UE and the application node needs to be transmitted through AN, UP Function, NG5, and Data Network. According to ETSI TS 123 401 [27], its end-to-end latency is about 20 ~ 200 ms, which obviously cannot meet URLLC requirements shown in Table 11. In order to reduce propagation time, the study [4] shows that the nodes providing applications can be distributed deployed close to the network edge with NFV enable technology. Then UE can select the nearest node when attaching mobile network.

Despite of nodes edge deployment, the movement of the UE will enlarge the propagation distance between UE and the node, thus increasing the end-to-end latency. In order to optimization latency, the simplest way is to always select the closest network edge node [6]. So these latency targets are required to migrate application instances (VM) or application-specific user-related context between different edge nodes. But by doing so, the least latency caused by application mobility is averagely above 15 ms [i.21], [i.30] and will cause a short period of interruption. So it is necessary for NGP to consider reducing the latency caused by application mobility.

8.14.2.3 E2E latency enlarged by Indeterminate processing delay in network nodes

To fulfil URLLC low latency requirement, much has been done to reduce propagation time by placing application node closer to UE with NFV. Even so, the packet transmission between UE and application node need go through physical or virtual network nodes, like routers or virtual switches and more. The processing delay from different node is indeterminate and remains an indispensable component of latency [i.31].

Now some solutions focus on reducing processing delay, DiffServ is a typical method. The DiffServ architecture provides Expedited Forwarding (IETF RFC 3246 [31]), so that low latency traffic can jump the queue of other traffic. However, when all traffic at one time is latency-sensitive, Then Diffserv is of little use [i.31]. In this situation, Diffserve architecture cannot ensure low processing delay. Now [i.31] and [i.32] are researching on how to make sure low processing delay all the time. NGP also should considerate how to make sure the deterministic processing delay to fulfil URLLC service.

8.14.2.4 Conflict between high reliability and low latency under wireless packet loss

Wireless packet loss (10e-3 [26]) is an important issue of concern. Packet loss leads to the failure of meeting the requirement of URLLC (10E-09 [25]). Reducing the packet loss is one of the issues concerned by NGP. Overcome packet loss by retransmission will enlarge end-to-end latency. So retransmission may not be a good idea for URLLC. It's important to reduce wireless packet loss not at the expense of increasing latency. Multi-path may be a good choice.

In addition, performance of transport protocol is affected by packet loss. For example, TCP throughput will be seriously decreased by packet loss. To reduce influence by wireless random packet loss, TCP could distinguish between random packet loss and congestion to do different response. So the cross-layer cooperation between mobile network and transmission network also needs to focus on.

In summary, in order to meet the requirements of URLLC, NGP needs to consider the characteristics of mobile network when conducting research on protocols.

8.14.3 Applicable Issues and Recommendations

Low Latency

Issue-01: Handover interruption results in an increment in end-to-end latency.

Recommendation-01: Make before handover-UE connects to target AN before disconnecting from source AN.

NOTE: Minimize latency of Access Networks and optimize RACH process, within scope of access networks e.g. 3GPP.

Issue-02: Interruption latency caused by mobility of application.

Recommendation-02: Enhance coordination between application and mobile network to keep seamless service continuity when migrate application instances (VM) or application-specific user-related information [25] between different edge nodes.

Issue-03: E2E latency enlarged by indeterminate processing delay in network nodes.

Recommendation-03: Ensure deterministic processing delay of the URLLC data stream in network nodes.

Ultra-Reliable

Issue-04: Wireless packet loss influence the performance of transport protocol.

Recommendation-04: NGP should support for transmission control accommodates the ultra-low latency requirements listed here for both transport and handover between access technologies and within the same access technology at the NGP architectural level.

8.14.4 Use cases

8.13.4.1 Case 1: Local UAV Collaboration

Unmanned aerial vehicles (UAVs) local vehicle collaboration can act as a mobile sensor network to autonomously execute sensing tasks in uncertain and dynamic environments while being controlled by a single user. Accuracy in sensing tasks is increased when deploying a team of UAVs versus just one as there are multiple vantage points using multiple sensors. Examples of uses for deploying a team of UAVs include:

- Searching for an intruder or suspect.
- Continual monitoring of natural disasters.
- Performing autonomous mapping.
- Collaborative manipulation of an object (e.g. picking up corners of a net or picking up a log).

The 3GPP system shall support:

- Latency of [10 ms] as collaboration requires vehicle altitude and position control loops to synchronize. Latency is required on the order of the control loop bandwidths.
- Near [100 %] reliability as instability and crashing of UAV could result from loss of communications. Control functions depend on this communication.
- Security to be provided at the level for current aviation Air Traffic Control (ATC) for command and control of vehicles in controlled airspace.
- Priority, Precedence, Preemption (PPP) needed as failure to transmit communications in reliable and timely manner could result in loss of property or life.
- Position accuracy within [10 cm] due to multiple UAVs that may need to collaborate in close proximity to one another.

8.14.4.2 Case 2: Industrial Factory Automation

Factory automation requires communications for closed-loop control applications. Examples for such applications are robot manufacturing, round-table production, machine tools, packaging and printing machines. In these applications, a controller interacts with large number of sensors and actuators (up to 300), typically confined to a rather small manufacturing unit (e.g. 10 m × 10 m × 3 m). The resulting S/A density is often very high (up to $1/m^3$). Many of such manufacturing units may have to be supported within close proximity within a factory (e.g. up to 100 in assembly line production, car industry).

In the closed-loop control application, the controller periodically submits instructions to a set of S/A devices, which return a response within a cycle time. The messages, referred to as telegrams, typically have small size (< 50 Bytes). The cycle time ranges between 2 and 20 ms setting stringent latency constraints on to telegram forwarding (< 1 ms to10 ms). Additional constraints on isochronous telegram delivery add tight constraints on jitter (10 - 100 us). Transport is also subject to stringent reliability requirements measured by the fraction of events where the cycle time could not be met (< 10e-9).

8.14.4.3 Case 3: V2X

In order to enhance the safety and reduce the fatigue during driving, assisted driving and autopilot are favored by more and more people. To ensure safety, communication between vehicle and vehicle as well as vehicle and application node is necessary in V2X. The transmission of such messages requires low latency and ultra reliability (a residual error rate of < 10e-9, latency < 1 - 15 ms) 3GPP TR 22.891 [i.1].

8.14.4.4 Scenario Targets

Table 12: URLLC Target KPIs

KPI Name	Description	Unites	Current behavior	Target Value/behavior
E2E latency	Time from APP to UE	ms	20 ~ 200 ms (Best effort)	< 10 ms (Deterministic)
handover latency	Time caused by handover	ms	> 45 ms	< 10 ms
Interruption latency during application migration	Time that application node out of service caused by application migration	ms	> 15 ms	< 10 ms
Reliability	The lose rate during transportation	%	10e-3	10e-9

Annex A (informative): Use Cases & Parameterization

This annex lists all of the Use Cases from the 3GPP TR 22.891 [i.1] SMARTER document and provides informative examples of typical feature and performance values for them.

Table A.1: Feature Parameterization of Use Cases

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
1a	Ultra-reliable comms	IoT	Industrial Control systems for a Utility (Gas Governor system across 10 km²) (Failsafe system management for 1 Hr).	Service	IoT (staticSA)	Reserved	Static	6Nines	4,5Nines	Outdoor	No
1b	Ultra-reliable comms	IoT	Mobile Health Monitoring (Failsafe local management for 1 Hr) Assume can turn on and off remote (minimum) HD video, compressed with AVC.		IoT (mobileConc)	Reserved	Low	7Nines	4,5Nines	Home	No
1c	Ultra-reliable comms	IoT	Real Time Vehicle Control w/remote control and local control intelligence w/soft management between local and remote control w/failover to local when remote comms fails and then C2C and C2R (to roadside) fail-safes takeover (Failsafe local management should operate for 5 Mins before remote mgt hands over to next BS) Assume can turn on and off remote (minimum) HD video, compressed with AVC compression.		IoT (mobileConc)	Reserved	High	7Nines	5Nines	Outdoor	No
2	Network Slicing	MANO	all	Capability	N/A	All	N/A	4,5Nines	4,5Nines	N/A	Yes
3a	Lifeline comms (natural disaster)	Public Safety	Basic Emergency Services Communications (Text, Voice (OTT or IMS) and basic rate data).	Service	Mobile-Tablet-Phablet	Dynamic Priority	Medium	6Nines	4,5Nines	Outdoor	Yes

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
3b	Lifeline comms (natural disaster)	Public Safety	D2D Emergency Services.		Mobile-Tablet-Phablet	Dynamic Priority	Low	6Nines	4,5Nines	Outdoor	Yes
3c	Lifeline comms (natural disaster)	Public Safety	D2D Public.		Mobile-Tablet-Phablet	Common	Low	5Nines	4,5Nines	Outdoor	Yes
4	Migration of Services from earlier generations	Legacy	Text, Voice (OTT or IMS) and basic rate data.	Capability	Mobile-Tablet-Phablet	Common	Medium	6Nines	4,5Nines	Outdoor	No
5	Mobile broadband for indoor scenario	MobileBB	Domestic User.	Service	Mobile-Tablet-Phablet	Common	Low	5Nines	4,5Nines	Home	No
6	Mobile broadband for hotspots scenario	MobileBB	Office Worker.	Service	Mobile-Tablet-Phablet	Dedicated	Low	5Nines	4,5Nines	Office	No
7	On-demand Networking	MobileBB	Nomadic, Event Based (social, updates).	Capability	Mobile-Tablet-Phablet	Dynamic Priority	Low	5Nines	4,5Nines	Outdoor	No
8	Use case for flexible application traffic routing	MANO	all	Capability	N/A	All	N/A	4,5Nines	4,5Nines	N/A	No
9	Flexibility and scalability	MANO	all	Capability	N/A	All	N/A	4,5Nines	4,5Nines	N/A	No
10	Mobile broadband services with seamless wide-area coverage	MobileBB	all	Capability	Mobile-Tablet-Phablet	Dedicated	Medium	5Nines	4,5Nines	Outdoor	No
11	Virtual Presence	New	360° Video-Conferencing.	Service	Mobile-Tablet-Phablet	Dynamic Priority	Low	5Nines	4,5Nines	Office	No

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
12	Connectivity for drones	New	High Speed Video imaging systems and (S&A) It is assumed that the unit camps on Macro sites and has the ability to reselect base stations so that more than one BS in the area can be seen and 2 or more BS would have to fail before the system would be self-flying It is also assumed that the drone flies at 30 - 70 mph but reduced to no more than 40 mph during comms loss and can avoid obstacles itself for more than 32 s at a time with remedial collision avoidance failsafe algorithms.	Service	IoT (mobileSA)	Dynamic Priority	High	7Nines	6Nines	Outdoor	Yes
13	Industrial Control	IoT	Smart Grid and industrial monitoring capability.	Service	IoT (mobileConc)	Reserved	Low	7Nines	4,5Nines	Factory	Yes
14	Tactile Internet	New	Remote operation of tools.	Service	Mobile-Tablet-Phablet	Dedicated	Low	7Nines	7Nines	Office	Yes
14	Tactile Internet	New	Remote operation of tools.		Mobile-Tablet-Phablet	Dedicated	Low	7Nines	7Nines	Medical	Yes
15	Localized real-time control	New	Control of a 1 or more robots from a local CC/CM connected to the rest of the 5G network (failsafe robots that can gracefully exit task if comms fails and move to stationary/standby).	Service	IoT (mobileConc)	Reserved	Low	7Nines	4,5Nines	Factory	Yes
16	Coexistence with legacy systems	Legacy	Ability to handover to LTE.	Capability	N/A	Not required	N/A	4,5Nines	4,5Nines	N/A	No
17	Extreme real-time comms & the tactile internet	New		Service	Mobile-Tablet-Phablet	Dedicated	Low	7Nines	7Nines	Office	Yes
18	Remote Control	New	High Speed Video imaging systems and (S&A) It is assumed that the unit camps on Macro sites and has the ability to reselect base stations so that more than one BS in the area can be seen and 2 or more BS would have to fail before the system would be self-flying It is also assumed that the system can failsafe and move to stationary/standby if required and reboot quickly on recovery. Assume down time of 32 s/yr is ok for comms if system is failsafe itself.	Service	IoT (mobileSA)	Dynamic Priority	High	7Nines	6Nines	Outdoor	Yes
19	Light weight device config	IoT	Light weight, remote (S and Config) configuration capability.	Capability	IoT (staticSA)	Common	Static	4,5Nines	4,5Nines	Outdoor	Yes

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
20	Wide area sensor monitoring and event driven alarms	IoT	Low cost, could be throwaway devices, may be 5G may be reporting via concentrator (analysis here is for concentrator which it is assumed may be mobile (e.g. flyover)).	Service	IoT (mobileConc)	Common	Medium	7Nines	4,5Nines	Outdoor	Yes
21	IoT Device Initialization	IoT	IoT device over-air IoT and Network provisioning and certification.	Capability	N/A	Common	N/A	4,5Nines	4,5Nines	N/A	Yes
22	Subscription security credentials update	IoT	IoT device subscription and security credentials update.	Capability	N/A	Common	N/A	4,5Nines	4,5Nines	N/A	Yes
23	Access from less trusted networks	Legacy	Evolution so that IMSI is ideally never sent over air unencrypted as transit networks can capture IMSI and identify user or spoof user. Need a method to stop this or protect identify when roaming.	Capability	N/A	Common	N/A	4,5Nines	4,5Nines	N/A	Yes
24	Bio-connectivity	New	Continuous and automatic medical telemetry of Blood pressure temp, etc.	Service	Wearables	Dynamic Priority	Medium	7Nines	4,5Nines	all	Yes
25	Wearable Device Communicatio n	New	Ability for wearable device to connect to a network via a stored phone on the user's person or when near to a mobile device. E.g. Smart Watch, Smart Running Shoe, etc.	Service	Wearables	Common	Medium	4,5Nines	4,5Nines	all	Yes
26	Best Connection per Traffic Type	New	Ability for the device to support multiple connections to multiple networks at once for various different traffic types according to service offered by connection type e.g. voice over Mobile and data over Wi-Fi/third party service.	Capability	Mobile-Tablet-Phablet	Common	N/A	4,5Nines	4,5Nines	N/A	No
27	Multi Access network integration	New	Ability to support multi-access network types at a time, e.g. V2V, IoT and other non-3GPP networks in a coordinated manner to support a multi-access network connection for a device or application.	Capability	Mobile-Tablet-Phablet	Common	N/A	4,5Nines	4,5Nines	all	Yes
28	Multiple RAT connectivity and RAT selection	New	e.g. 5G-RF, 5G-mm-Wave and LTE and multiplexing between to avoid handovers and maximize available connectivity at a time.	Capability	Mobile-Tablet-Phablet	All	N/A	4,5Nines	4,5Nines	all	Yes
29	Higher User Mobility	Environ	HST(High Speed Train) and Airplane connectivity.	Capability	Mobile-Tablet-Phablet	All	HST	4,5Nines	4,5Nines	HST	No
30	Connectivity Everywhere	Environ	Commercial and recreational UAV will provide Mobile Broadband, Cruising Ships also and in-flight connectivity.	Capability	N/A	All	N/A	4,5Nines	4,5Nines	all	No

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
31	Temporary Service for Users of Other Operators in Emergency Case	Public Safety	To provide High Availability connectivity during emergencies a subscriber of one network should be able to use a non-subscriber other network in its vicinity to make an emergency call.	Capability	N/A	Not required	N/A	7Nines	7Nines	all	Yes
32	Improvement of network capabilities for vehicular case	Environ	High rate services in and between cars, either as in car base stations and/or via a connected mobile device in the car.	Service	AV	Not required	High	7Nines	7Nines	Outdoor	Yes
33	Connected vehicles	Environ	Autonomous Vehicle (auto-drive) 1 ms response time over air and near 100 % reliability and coverage on all major roads.	Service	AV	Not required	High	7Nines	7Nines	Outdoor	Yes
34	Mobility on demand	Environ	Ability for user to select high mobility/low mobility/nomadic or static mobility service option.	Capability	Mobile-Tablet-Phablet	Not required	all	4,5Nines	4,5Nines	all	No
35	Context Awareness to support network elasticity	Environ	This is the CA driven, ANO use case where User Profile information can be collected securely and privately to enable the network to support rapid network reconfiguration to support user changing traffic and mobility load patterns.	Capability	Mobile-Tablet-Phablet	All	all	4,5Nines	4,5Nines	all	Yes
36	In-network caching	MobileBB	The ability to cache information in the network to speed up downloads of common information and reduce load over the network to internet transmission.	Capability	N/A	Not required	N/A	4,5Nines	4,5Nines	all	Yes
37	Routing path optimization when server changes	MobileBB	Optimal routing and server selection for virtual nailed up streams for highly available/fast temporal or permanent stream demands such as Immersive Video/Tactile internet or ad-hoc broadcast info.	Capability	N/A	Not required	N/A	4,5Nines	4,5Nines	N/A	No
38	ICN Content Retrieval	MobileBB	Adoption of control for caching, routing and discovery of content.	Capability	Mobile-Tablet-Phablet	Not required	N/A	4,5Nines	4,5Nines	N/A	Yes

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
39	Wireless Briefcase	New	This use case provides a user with Personal Content Management (PeCM) of all of their traditionally stored HDD information in the form of a Flat Distributed Personal Cloud (FDPeC) facilitated over the 3GPP communications network. 5GIC: Personal Content Management "Wireless Briefcase" to support NGMN Smart office See Mobile Broadband cases, plus additional control for indexing Most Recently Used(MRU)/Least Recently Used(LRU) documents/files on their device to maintain a Distributed HDD Store in network (Distributed Personal Cloud) (DPC).	Capability	Mobile-Tablet-Phablet	DynamicPriority	Low	6Nines	6Nines	all	Yes
40	Devices with variable data	IoT	Ability to deploy a device that can be operated on an ad-hoc basis to perform some simple electronic task and relay its information collected back to a different physical point. E.g. a landslide sensor triggered device that relays a picture when a land slide is triggered and sends a video image back to a central point to assess the damage.	Capability	IoT(staticS)	Not required	N/A	4,5Nines	4,5Nines	Outdoor	Yes
41	Domestic Home Monitoring	IoT	IoT Concentrator device for home monitoring usage. 5GIC notes: this is a combination of other IoT cases using non-5G capillary devices connected to 5G-Mobile as concentrator.	Service	IoT(mobileSA)	Not required	Static	4,5Nines	4,5Nines	Home	Yes
42	Low mobility devices	IoT	e.g. support for static or near static devices such as IoT sensors on a bridge sensing stress/strain and reporting back to a maintenance unit over network.	Capability	IoT(mobileSA)	Not required	Low	4,5Nines	4,5Nines	Outdoor	Yes
43	Materials and inventory management and location tracking	IoT	Support of IoT tagging of warehouse/stores materials and equipment and interworking with warehouse/store management equipment such as vehicles in the building or conveyors, etc.	Service	IoT(mobileS)	Not required	Static	4,5Nines	4,5Nines	all	Yes
44	Cloud Robotics	IoT	Communication support for relatively dumb Robots and remote connected and controlling cloud intelligence that drives them.	Service	UAV	Dynamic Priority	High	7Nines	7Nines	all	Yes

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
45	Industrial Factory Automation	IoT	Communications support for closed loop factory control systems.	Service	IoT(mobileSA)	Dynamic Priority	Static	7Nines	7Nines	Factory	Yes
46	Industrial Process Automation	IoT	Communications support for open loop and supervisory systems for factory control systems.	Service	IoT(mobileSA)	Dynamic Priority	Static	6Nines	6Nines	Factory	Yes
47	SMARTER Service Continuity	Environment	Maximizes the IP routing part of the ETE connection and minimizes the GTP tunnelled part of the network. Note HTTP and DASH can both accommodate a change in IP endpoint address as they are URL based not IP endpoint based protocols.	Capability	N/A	Not required	N/A	4,5Nines	4,5Nines	N/A	No
48	Provision of essential services for very low-ARPU areas	Legacy	Remote service provision with coverage throughput but slightly longer latency on the MBB link.	Service	Mobile-Tablet-Phablet	Not required	Low	4,5Nines	4,5Nines	Outdoor	Yes
49	Network capability exposure	Environment	E.g. Exposing the network Slicing capability to third parties as well as services like OTT voice in coordination with the Network Operator for lower cost third party service offerings.	Capability	N/A	All	N/A	4,5Nines	4,5Nines	N/A	No
50	Low-delay speech coding	Legacy	Ability to reduce speech coding delays for applications such as gaming with live voice from 20 ms - 40 ms today towards only 10 ms one way in 5G timeframe.	Service	Mobile-Tablet-Phablet	Dynamic Priority	N/A	5Nines	5Nines	all	No
51	Network enhancement to support scalability and automation	Environment	Support of network scalability and offload ability during heavy load periods, provides network level load balancing and offload of lower QoS services to legacy and third party partners whilst retaining high QoS services such as UHD Video and VoLTE or Vo5G.	Capability	N/A	All	N/A	4,5Nines	4,5Nines	N/A	No
52	Wireless Self-Backhauling	Environment	E.g. using mm-Wave backhaul to backhaul the mm/RF cell coverage traffic at the same site thus avoiding wired backhaul.	Capability	N/A	Not required	N/A	4,5Nines	4,5Nines	N/A	No
53	Vehicular Internet & Infotainment	MobileBB	Communications provision of internet by mobile network for the dedicated purpose of infotainment in the car or vehicle where this comms device is left in the car as the de-facto infotainment device for the car rather than DAB, FM-Radio, DVD/CD or Mobile Device that are common-place today.	Service	AV	Not required	High	4,5Nines	4,5Nines	Outdoor	Yes

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
54	Local UAV Collaboration	IoT	Local UAV collaboration in a community where comms network is used to control the UAV and report back to the connected mobile devices subscribed. E.g. Burglar monitoring, local sensor network, etc.	Service	UAV	Dynamic Priority	High	7Nines	7Nines	Outdoor	Yes
55	High Accuracy Enhanced Positioning	Environment	ePositioning for 95 % of service are to 1 m accuracy.	Capability	Mobile-Tablet-Phablet	Not required	all	4,5Nines	4,5Nines	All	Yes
56	Broadcasting Support	MBMS	The system should be able to support an enhanced form of MBMS that includes transmission of scheduled linear time Audio and Audio &Video programmes.	Service	Mobile-Tablet-Phablet	Dedicated	N/A	4,5Nines	4,5Nines	All	Yes
57	Ad-Hoc Broadcasting	MBMS	MBMS and eMBMS are defined in UMTS and LTE. Take-up has been poor. However, there is a demand for good quality event based content broadcasting over and above IP web pages and video snippets. This use case proposes the ability to setup event based video content broadcasting, using a slice of the local or temporary 3GPP system in the environ of the event.	Service	Mobile-Tablet-Phablet	Dynamic Priority	Medium	5Nines	4,5Nines	Outdoor	Yes
58	Use Case for Green Radio	Environment	Reduction in Mobile systems energy consumption by 100 times as compared to 4G.	Capability	N/A	Not required	N/A	N/A	N/A	All	No
59	Massive Internet-Of-Things M2M and device identification	IoT	Support of up to 200 000 sensors/km² ITU-T or more realistically 5G-PPP mentions a device density of 1 M/km².	Service	IoT(staticS)	Not required	all	4,5Nines	4,5Nines	All	Yes
60	Light weight device communication	IoT	Low cost, MTC remote IoT, non-IP connected devices.	Service	IoT(staticSA)	Common	Static	4,5Nines	4,5Nines	All	No
61	Fronthaul/ Backhaul Network Sharing	Environment	Ability to share Fronthaul and/or Backhaul across networks.	Capability	N/A	All	N/A	5Nines	5Nines	All	Yes
62	Device Theft Preventions/ Stolen Device Recovery	Environment	Service operated across access device and network.	Service	All	All	N/A	5Nines	5Nines	All	Yes

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
63	Diversified Connectivity	Environment	Support of devices that include some that may belong to one entity at a time and some that may belong to one entity at one time and another at another time. The entity may be a building, a person, a vehicle or a machine/thing.	Capability	All	N/A	N/A	5Nines	5Nines	All	Yes
64	User Multi-Connectivity across operators	Environment	Enables devices or cluster of devices to provide connectivity to 2 or more networks at a time where the networks may be providing some common services but also some different services.	Capability	All	N/A	Medium	5Nines	5Nines	All	Yes
65	Moving ambulance and bio-connectivity	New	Provision of services to support ambulances and biodiversity feeds.	Service	Medical devices	Dedicated	Medium	6Nines	6Nines	All	Yes
66	Broadband Direct Air to Ground Communications (DA2GC)	eMBB	Voice video, telephony data services to aircraft.	Capability	Mobile-Tablet-Phablet	N/A	High	5Nines	5Nines	Flight	Yes
67	Wearable Device Charging	eMBB	Charging for wearable devices.	Capability	All	N/A	All	5Nines	5Nines	All	No
68	Telemedicine Support	New	Provision of services to support ambulances and biodiversity feeds.	Service	Medical Devices, Mobile-Tablet-Phablet	Dedicated	Medium	6Nines	6Nines	All	Yes
69	Network Slicing - Roaming	New	Roaming for network slicing.	Capability	All	All	All	6Nines	6Nines	All	Yes
70	Broadcast/ Multicast Services using a Dedicated Radio Carrier	New	Broadcast over dedicated resources.	Capability	All	Dynamic Priority	All	5Nines	5Nines	All	No
71	Wireless Local Loop	eMBB	Wireless local loop.	Capability	All	Dynamic Priority	All	5Nines	5Nines	All	No
72	5G Connectivity Using Satellites	eMBB	Satellite connectivity.	Capability	All	Dynamic Priority	All	5Nines	5Nines	All	No

Ref	Use Case	Use Case Group	Example System/UC notes	Capability, Service	Device Type (=DeviceType +UsageProfile)	Slicing Reserved Dedicated Dyn.Priority Common	Mobility	System Reliability	5G Reliability	Locale (local location type)	Security Impact (Additional)
73	Delivery Assurance for High Latency Tolerant Services		Highly reliable but delay tolerant services, should have confidence of delivery of information.	Capability	All	N/A	Static to Medium	6Nines	6Nines	All	Yes
74	Priority, QoS and Policy Control		Network provision of priority, QoS and policy based service management.	Capability	All	All	All	6Nines	6Nines	All	Yes

Table A.2: Performance Parameterization of Use Cases

Ref	Use Case	Typical Rb/user 1xUsr/Cell (50 %) (IP-Mbit/s)	Typical Rb/user, 1xUsr/Cell (95 %) (IP-Mbit/s)	Typical Rb/user Loaded Cell (50 %) (IP-Mbit/s)	Typical Rb/user, Loaded Cell (95 %) (IP-Mbit/s)	Connection/Session/Call Setup (ms)	Connection Rejoin	Latency Device to Nwk Store (Pk/Msg) (ms) (MEC BS)	Latency Device to Web Edge (Pk/Msg) (ms) (MEC CN)	Connection Type Required for UP transport	Access Type (s) Required
1a	Ultra-reliable comms	1	0,3	0,2	0,1	5	1	2	5	Group	Rfcellular
1b	Ultra-reliable comms	100	50	30	15	5	1	2	5	Group	All
1c	Ultra-reliable comms	100	50	30	15	5	1	2	5	Group	All
2	Network Slicing										
3	Lifeline comms (natural disaster)	1	0,3	0,2	0,1	200	20	50	100	Group	Rfcellular
4	Migration of Services from earlier generations	1	0,3	0,2	0,1	300	100	50	100	Group	Rfcellular
5	Mobile broadband for indoor scenario	10 000	2 000	1 000	500	100	20	5	10	Dedicated	All
6	Mobile broadband for hotspots scenario	10 000	2 000	1 000	500	100	20	5	10	Dedicated	All
7	On-demand Networking	5 000	1 000	500	150	100	20	5	10	Group	All

Ref	Use Case	Typical Rb/user 1xUsr/Cell (50 %) (IP-Mbit/s)	Typical Rb/user, 1xUsr/Cell (95 %) (IP-Mbit/s)	Typical Rb/user Loaded Cell (50 %) (IP-Mbit/s)	Typical Rb/user, Loaded Cell (95 %) (IP-Mbit/s)	Connection/Session/Call Setup (ms)	Connection Rejoin	Latency Device to Nwk Store (Pk/Msg) (ms) (MEC BS)	Latency Device to Web Edge (Pk/Msg) (ms) (MEC CN)	Connection Type Required for UP transport	Access Type (s) Required
9	Flexibility and scalability	Capability UC needs to be paired with another Service UC									
10	Mobile broadband services with seamless wide-area coverage	10 000	2 000	1 000	500	100	20	10	5	Dedicated	All
11	Virtual Presence	10 000	2 000	1 000	500	100	N/A	10	5	Dedicated	All
12	Connectivity for drones	1 000	500	200	100	100	N/A	10	5	Dedicated	All
13	Industrial Control	100	50	30	15	5	1	5	2	Group	All
14	Tactile Internet	10 000	2 000	1 000	500	100	1	5	2	Dedicated	All
15	Localized real-time control	1 000	500	200	100	100	N/A	10	5	Group	All
16	Coexistence with legacy systems	Capability UC needs to be paired with another Service UC									
17	Extreme real-time comms & the tactile internet	10 000	2 000	1 000	500	100	1	5	2	Dedicated	All
18	Remote Control	1 000	500	200	100	100	N/A	10	5	Dedicated	All
19	Light weight device config	1	0,3	0,2	0,1	100	20	50	100	Group	Rfcellular
20	Wide area sensor monitoring and event driven alarms	1	0,3	0,2	0,1	100	20	50	100	Group	Rfcellular
21	IoT Device Initialization	1	0,3	0,2	0,1	300	N/A	50	100	Group	All
22	Subscription security credentials update	1	0,3	0,2	0,1	300	N/A	50	100	Group	All
23	Access from less trusted networks	Capability UC needs to be paired with another Service UC									
24	Bio-connectivity	1	0,3	0,2	0,1	300	20	50	25	Group	All

Ref	Use Case	Typical Rb/user 1xUsr/Cell (50 %) (IP-Mbit/s)	Typical Rb/user, 1xUsr/Cell (95 %) (IP-Mbit/s)	Typical Rb/user Loaded Cell (50 %) (IP-Mbit/s)	Typical Rb/user, Loaded Cell (95 %) (IP-Mbit/s)	Connection/Session/Call Setup (ms)	Connection Rejoin	Latency Device to Nwk Store (Pk/Msg) (ms) (MEC BS)	Latency Device to Web Edge (Pk/Msg) (ms) (MEC CN)	Connection Type Required for UP transport	Access Type (s) Required
25	Wearable Device Communication	1	0,3	0,2	0,1	300	N/A	50	100	Group	All
26	Best Connection per Traffic Type	Capability UC needs to be paired with another Service UC									
27	Multi Access network integration	Capability UC needs to be paired with another Service UC									
28	Multiple RAT connectivity and RAT selection	Capability UC needs to be paired with another Service UC									
29	Higher User Mobility	100	50	30	15	10	5	10	5	Dedicated	RF(Cellular)

Ref	Use Case	Typical Rb/user 1xUsr/Cell (50 %) (IP-Mbit/s)	Typical Rb/user, 1xUsr/Cell (95 %) (IP-Mbit/s)	Typical Rb/user Loaded Cell (50 %) (IP-Mbit/s)	Typical Rb/user, Loaded Cell (95 %) (IP-Mbit/s)	Connection/Session/Call Setup (ms)	Connection Rejoin	Latency Device to Nwk Store (Pk/Msg) (ms) (MEC BS)	Latency Device to Web Edge (Pk/Msg) (ms) (MEC CN)	Connection Type Required for UP transport	Access Type (s) Required
30	Connectivity Everywhere	Capability UC needs to be paired with another Service UC									
31	Temporary Service for Users of Other Operators in Emergency Case	Capability UC needs to be paired with another Service UC									
32	Improvement of network capabilities for vehicular case	1 000	500	200	100	100	N/A	10	5	Dedicated	All
33	Connected vehicles	100	50	30	15	5	1	5	2	Group	All
34	Mobility on demand	Capability UC needs to be paired with another Service UC									
35	Context Awareness to support network elasticity	Capability UC needs to be paired with another Service UC									
36	In-network caching	Capability UC needs to be paired with another Service UC									
37	Routing path optimization when server changes	Capability UC needs to be paired with another Service UC									
38	ICN Based Content Retrieval	Capability UC needs to be paired with another Service UC									
39	Wireless Briefcase	Capability UC needs to be paired with another Service UC									
40	Devices with variable data	Capability UC needs to be paired with another Service UC									
41	Domestic Home Monitoring	1	0,3	0,2	0,1	300	20	50	25	Group	All
42	Low mobility devices	Capability UC needs to be paired with another Service UC									

Ref	Use Case	Typical Rb/user 1xUsr/Cell (50 %) (IP-Mbit/s)	Typical Rb/user, 1xUsr/Cell (95 %) (IP-Mbit/s)	Typical Rb/user Loaded Cell (50 %) (IP-Mbit/s)	Typical Rb/user, Loaded Cell (95 %) (IP-Mbit/s)	Connection/Session/Call Setup (ms)	Connection Rejoin	Latency Device to Nwk Store (Pk/Msg) (ms) (MEC BS)	Latency Device to Web Edge (Pk/Msg) (ms) (MEC CN)	Connection Type Required for UP transport	Access Type (s) Required
43	Materials and inventory management and location tracking	1	0,3	0,2	0,1	300	20	50	25	Group	All
44	Cloud Robotics	No reference values available, depends on Robot type, speed etc.									
45	Industrial Factory Automation	100	50	30	15	5	1	5	2	Group	All
46	Industrial Process Automation	100	50	30	15	5	1	5	2	Group	All
47	SMARTER Service Continuity	Capability UC needs to be paired with another Service UC									
48	Provision of essential services for very low-ARPU areas	50	20	20	10	5	1	5	2	Group	All
49	Network capability exposure	Capability UC needs to be paired with another Service UC									
50	Low-delay speech coding	1	0,3	0,2	0,1	50	10	N/A	N/A	Dedicated	All
51	Network enhancements to support scalability and automation	Capability UC needs to be paired with another Service UC									
52	Wireless Self-Backhauling	Capability UC needs to be paired with another Service UC									
53	Vehicular Internet & Infotainment	500	200	100	75	150	N/A	10	5	Dedicated	All
54	Local UAV Collaboration	Capability UC needs to be paired with another Service UC									
55	High Accuracy Enhanced Positioning (ePositioning)	Capability UC needs to be paired with another Service UC									

Ref	Use Case	Typical Rb/user 1xUsr/Cell (50 %) (IP-Mbit/s)	Typical Rb/user, 1xUsr/Cell (95 %) (IP-Mbit/s)	Typical Rb/user Loaded Cell (50 %) (IP-Mbit/s)	Typical Rb/user, Loaded Cell (95 %) (IP-Mbit/s)	Connection/Session/Call Setup (ms)	Connection Rejoin	Latency Device to Nwk Store (Pk/Msg) (ms) (MEC BS)	Latency Device to Web Edge (Pk/Msg) (ms) (MEC CN)	Connection Type Required for UP transport	Access Type (s) Required
56	Broadcasting Support	500	200	100	300	100	20	100	50	Group	All
57	Ad-Hoc Broadcasting	500	200	100	300	100	20	100	50	Group	All
58	Use Case for Green Radio	Capability UC needs to be paired with another Service UC									
59	Massive Internet-Of-Things M2M and device identification	1	0,3	0,2	0,1	300	20	50	25	Group	All
60-74	Not updated as a lot of the performance criteria are covered by other Use cases And most are capabilities rather than performance impacting UC										

Annex B (informative): 5G mobile network model

B.0 Introduction

This annex provides a reference model for the 3GPP Release 15 5G Next Generation Core Network (NGCN) and the differences from the 4G/LTE Evolved Packet System.

B.1 References

The NGCN is defined in ETSI TS 123 501 [32], with the two architecture representations (point-to-point and service based) specified in 3GPP TR 23.799 [i.2].

B.2 Reading the scenarios in the context of 5G

Cellular scenarios in earlier versions of the present document were written and illustrated using LTE ('4G') network terminology; in particular clause 8.1, "Addressing" and clause 8.3, "Mobility". Clause B.4 maps this LTE terminology, specifically network function names and acronyms, to the equivalent 5G terminology.

B.3 Key differences from LTE mobile Network Model

1) A Service Based Architecture (see clause B.4).

2) Amended protocol stacks for control plane and user plane (see clause B.5).

3) Network Slicing support. This allows network resources to be allocated and configured together, supporting the ad-hoc creation and removal of logical networks to meet particular networking use cases - such as a temporary high-bandwidth or low-latency network slice.

4) Control and User Plane Separation (CUPS) is a native part of the 5G architecture. This allows independent scaling of each plane as appropriate.

B.4 Service Based Architecture

The 5G SBA is the intended deployment architecture for 5G. It allows network functions to discover each other and communicate control plane messages via HTTP APIs.

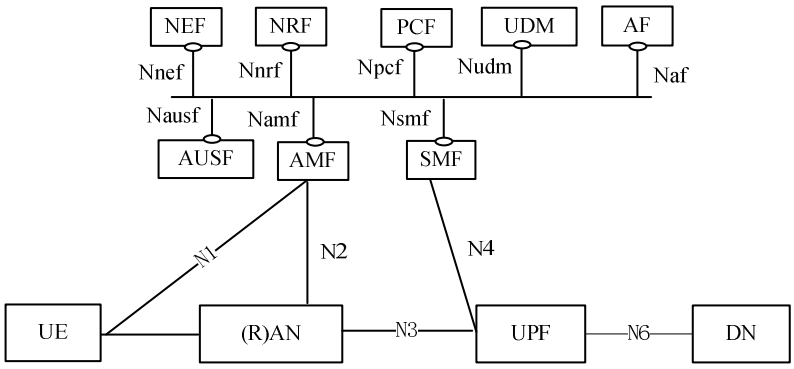


Figure B.1: 3GPP 5G Service Based Architecture

In addition to the point-to-point Network Functions, SBA includes Network Exposure Function (NEF), NF Repository Function (NRF). These are used to allow Network Function to register their own API endpoints, and discover the API endpoints of other NFs.

3GPP also define a point-to-point architecture as a reference representation, which is not intended to be deployed.

It shows the interactions between network functions, and is useful to compare the architecture to LTE:

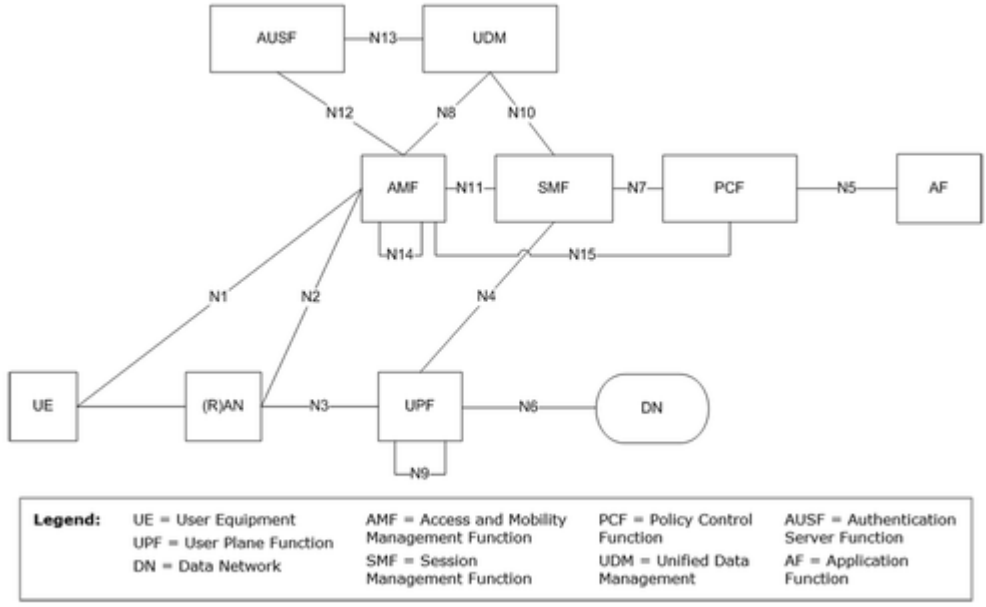


Figure B.2: 3GPP 5G point-to-point architecture representation

Many NFs (Network Functions) have been renamed or combined roles:

- User Plane Function maps to the LTE PGW/ SGW user planes
- Access and Mobility Management Function maps to the LTE MME
- Session Management Function maps to LTE MME, PGW/SGW control planes
- Authentication Server Function and User Data Management map to LTE HSS/AAA
- Policy Control Function maps to LTE PCRF

B.5 5G Protocol stacks

B.5.0 Introduction

This clause describes the user plane and control plane protocol stacks defined for 5G.

B.5.1 User plane protocols

For 5G phase 1 (Release 15): to summarize clause 11.2 of 3GPP TR 29.891 [33]:

- GTPv1-U will be used as the 5G UP Encapsulation protocol over the Access Network to NGCN interface (N3) and NGCN to external UPF (N9). (See ETSI TS 138 300 [34]).
- A 5GS Container will be carried within a GTP Extension Header over the N3 and N9 interfaces, to carry the information required to be sent in the encapsulation header over the N3 and N9 interfaces. 3GPP RAN3 will specify the contents of the 5GS Container.

- It is FFS whether CT4 needs to specify new GTP-U extension header(s) in ETSI TS 129 281 [35] for the 5GS Container.

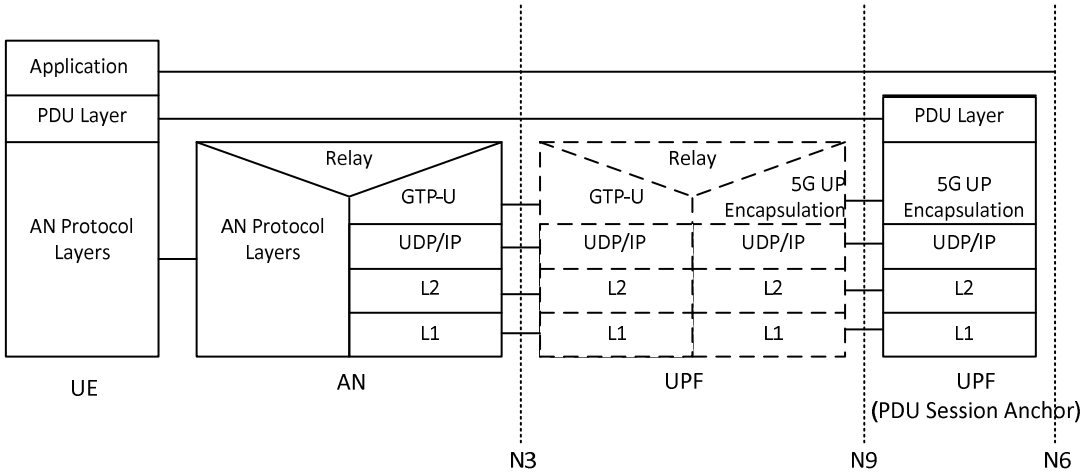
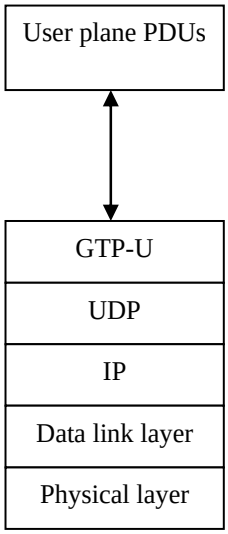


Figure B.3: 3GPP 5G Rel-15 UP Protocols



Legend:

- **PDU layer:** This layer corresponds to the PDU carried between the UE and the DN over the PDU session. When the PDU session Type is IPV6, it corresponds to IPv6 packets; When the PDU session Type is Ethernet, it corresponds to Ethernet frames; etc.
- **GPRS Tunnelling Protocol for the user plane (GTP-U):** This protocol supports multiplexing traffic of different PDU sessions (possibly corresponding to different PDU session Types) by tunnelling user data over N3 (i.e. between the AN node and the UPF) in the backbone network. GTP shall encapsulate all end user PDUs. It provides encapsulation on a per PDU session level. This layer carries also the marking associated with a QoS Flow.
- **5G Encapsulation:** This layer supports multiplexing traffic of different PDU sessions (possibly corresponding to different PDU session Types) over N9 (i.e. between different UPF of the 5GC). It provides encapsulation on a per PDU session level. This layer carries also the marking associated with a QoS Flow.

Figure B.4: User Plane Protocol Stack

B.5.2 5G Control Plane Protocols

Per clause 6.2 of 3GPP TR 29.891 [33], Release 15, the Service Based Architecture will utilize a REST-based API to replace the DIAMETER interfaces from previous releases. More recent IETF protocols, including QUIC (to replace TLS/TCP) and CBOR (Concise Object Binary Representation, to replace JSON) are under consideration for Release 16.

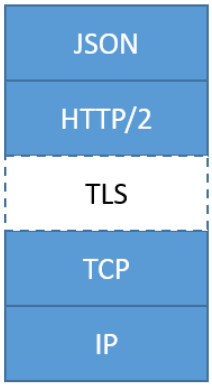


Figure B.5: 3GPP 5G Rel-15 CP Protocols

B.6 Networking issues carried over from LTE architecture

The following concerns raised in clause 6.1 (LTE Model) persist in the 5G architecture:

1) The use of GTP tunnels for multiplexing per-user data, adding to header size and processing overheads.

2) The reliance on interface-centric IP, resulting in a signalling storm during mobility as old tunnels are torn down and new tunnels set up.

3) Inefficient transmission of IP headers, or the use of compute-intensive header compression and decompression.

4) The lack of intrinsic security between radio access and the core network, requiring an IPsec encapsulation.

5) No interworking between the radio-scheduler's throughput control, and L4 end-to-end congestion controls, resulting in: inefficient use of available capacity, congestion caused by aggressive endpoints, wasteful retransmission at L4 for data queued at L2/L1.

Annex C (informative): Void

History

Document history		
V1.1.1	July 2024	Publication