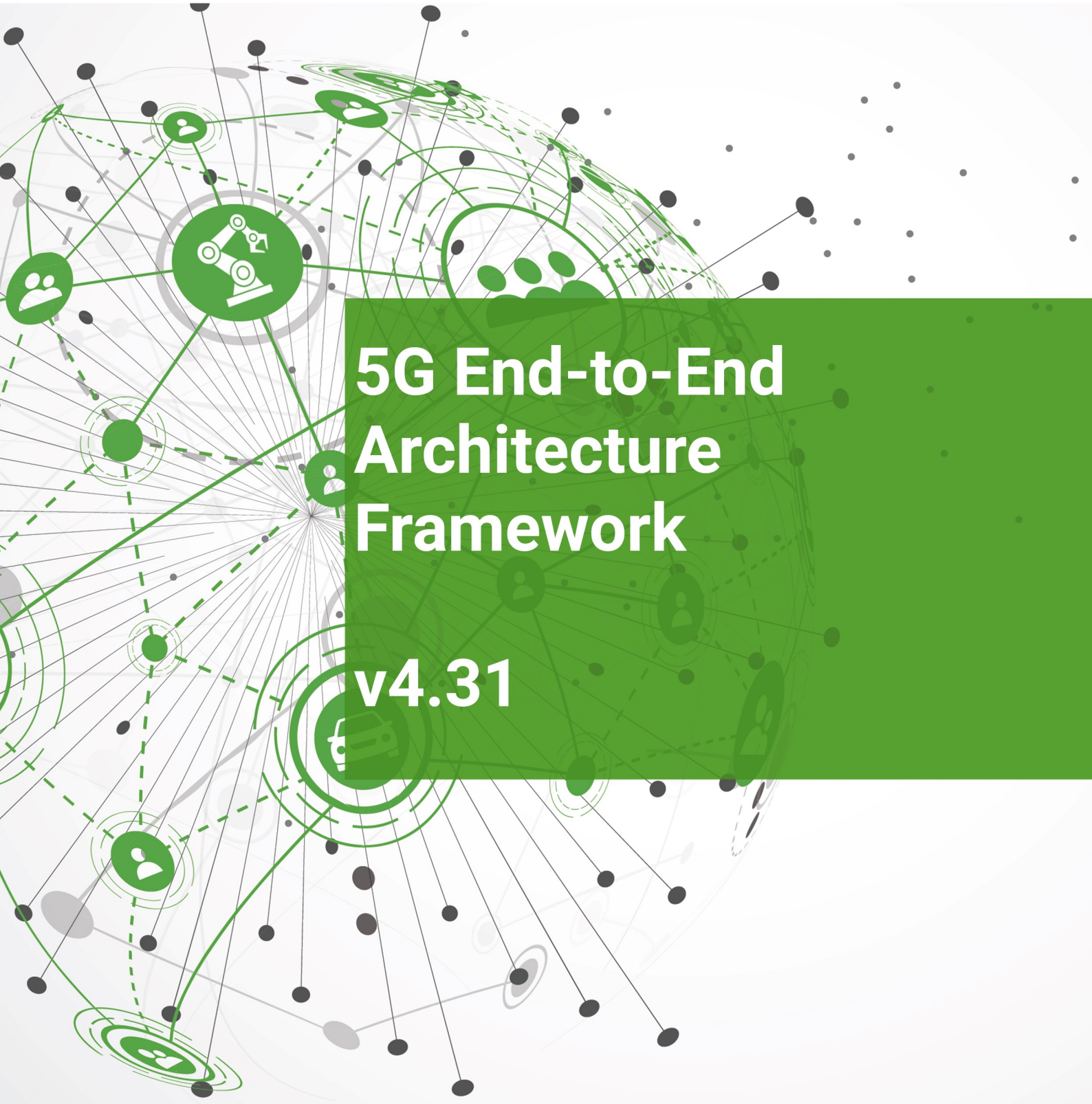
ngmn
the engine of
wireless innovation



5G End-to-End Architecture Framework

v4.31

5G End-to-End Architecture Framework

(Phase-3)

by NGMN Alliance

Version:	**V4.31**
Date:	**12-November-2020**
Document Type:	**Final Deliverable (approved)**
Confidentiality Class:	**P - Public**

Project:	**P1-Requirements and Architecture**
Editor / Submitter:	**Sebastian Thalanany**
Contributors:	**Sebastian Thalanany (U.S. Cellular), Srisakul Thakolsri (NTT DOCOMO) Tayeb Benmeriem (Orange), Gary Li (Intel), Marie-Paule Odini (HP), Fran O'Brien (Cisco), Sheeba Backia Mary B (Lenovo, Motorola Mobility), Andreas Kunz (Lenovo, Motorola Mobility), Minpeng (CMCC), Colin Blanchard (BT), Stan Wong (HKT Limited), Nicol So (CommScope), Farooq Bari (AT&T)**
Approved by / Date:	**NGMN Board, 10th November 2020**

Abstract: Short introduction and purpose of document

This document delineates the requirements in terms of entities and functions that characterise the capabilities of an E2E (end-to-end) framework. Architectural perspectives and considerations associated with the service categories - eMBB, mIoT, URLLC - envisioned for 5G (Fifth Generation) underscore the delineation of the E2E framework requirements. These requirements are intended as guidance in the development of inter-operable and market enabling specifications for a continuing advancement of the 5G ecosystem of heterogenous access, virtualization, autonomic capabilities, forward-looking service enablers, and emerging usage scenarios.

Table of Contents

1 Introduction 5
2 References 5
3 Definitions 8
4 Autonomic Networking 10
4.1 General 10
4.2 High level architecture 11
4.2.1 User Plane 15
4.2.2 Control Plane 15
4.2.3 Data Plane 15
4.2.4 Fulfillment Plane 15
4.2.5 Assurance Plane 15
4.2.6 Exposure Plane 15
4.2.7 Service Orchestration 16
4.2.8 Service Assurance 17
4.3 Closed-loop end-to-end system 19
4.3.1 Generic feedback model 19
4.3.2 Feedback control for autonomic behaviour 20
4.4 Usage scenarios 21
5 Network Data layer 22
5.1 General 22
5.2 Architectural considerations 22
5.3 Stateful and Stateless Considerations 23
5.4 Data Placement and Affinity Rules 23
5.5 Usage Scenarios 24
6 Artificial intelligence (AI) and machine learning (ML) 24
6.1 General 24
6.2 Broad categories of AI and ML 25
6.2.1 Supervised Learning 27
6.2.1.1 Examples of applicability 27
6.2.2 Unsupervised Learning 27
6.2.2.1 Examples of applicability 27
6.2.3 Reinforcement Learning 27
6.2.3.1 Examples of applicability 27
6.2.4 Federated Learning 27
6.2.4.1 Examples of applicability 29
6.2.4.1.1 Network optimization 29
6.2.4.1.2 Data privacy across different and cooperating/interacting entities 29
6.2.4.1.3 Application within the UN Sustainable Development Solutions Network (UN SDSN) 29
6.2.4.1.4 Multi-access Edge cloud performance enhancement 30
6.3 Architectural enhancements through cognitive capabilities 30
6.4 Context for AI and ML Data Model 31
6.4.1 Consumers 32
6.4.2 Core Network 32
6.4.3 Edge Network 32
6.4.4 Access Network 32
6.4.5 Transport Network 32
6.4.6 Platform as a Service 33
6.5 Guidance on Test and Certification of autonomics 33
6.5.1 Generic framework for testing and certifying autonomic functions 33
6.6 Usage Scenarios 35

6.6.1 Autonomic Networking 35
6.6.1.1 Fault detection in the FTTH network 35
6.6.1.2 Service Based Architecture 36
6.6.1.3 Network Slice Service Assurance 36
7 Virtualization in the radio access network 37
7.1 General 37
7.2 Architectural Considerations for Split Centralization and Distribution 37
7.2.1 Trade-offs associated with different split arrangements 38
7.2.1.1 High fronthaul bandwidth 39
7.2.1.2 Different Functional Split options: Centralization versus Distribution 39
7.2.1.3 Integration and operation complexity 39
7.2.1.4 Open and interoperable interfaces 39
7.3 Usage Scenarios 40
7.3.1 Network Resource Sharing among multiple tenants 40
7.3.2 Spectrum Sharing among multiple tenants 40
8 Multi-Access Edge Computing 41
8.1 General 41
8.2 Converged access and virtualization 41
8.3 Usage Scenarios 41
9 Distributed Ledger Technology 43
9.1 General 43
9.2 Smart Contract 44
9.3 DLT and MEC 44
9.4 DLT and Autonomic Networking 45
9.5 Usage Scenarios 45
9.5.1 Zero CAPEX model 45
9.5.2 DLT based Common Marketplace Platform 46
9.5.3 Exemplification of interactions in a Common Marketplace platform 49
9.5.4 Federated DLT platform in the Marketplace 50
10 Vertical Market 52
10.1 General 52
10.2 Ontology of identities and roles 52
10.3 Architectural considerations for a Vertical Market 53
11 End-to-End Security 55
11.1 General 55
11.2 Autonomic networking 55
11.3 Network Data Layer 55
11.4 AI and ML 57
11.5 Virtualization in the RAN 57
11.6 Multi-access Edge Computing 58
11.7 Distributed Ledger Technology 59
11.8 Vertical Market 59
11.9 Privacy 60
12 List of Abbreviations 61

1 INTRODUCTION

The purpose of this document is to provide a high-level framework of architectural principles and requirements that provide guidance and direction for NGMN partners and standards development organisations in the shaping of the 5G suite of interoperable capabilities, enablers, and services. It builds on the architectural concepts and directions described in the NGMN White Paper [1] [20], and related NGMN publications [2] [3]. This document describes advancements beyond the previous versions in terms of requirements, refinements, forward-looking technologies, and use cases that are relevant in an evolving 5G ecosystem.

The synergies across the emerging enablers is elaborated to reveal insights on forward-looking usage scenarios that inspire flexible business models to suit diverse deployment arrangements of access and services, from an end-to-end perspective. Along these directions, cognitive and self-organizing, system-wide ingredients are pivotal for an emerging and expansive ecosystem of services, and service experience advancement, together with open and interoperable revenue sharing architectural arrangements. These aspects are examined through a system-wide lens in terms of realizing the promise of a forward-looking service paradigm, enabled through a service-based framework of virtualization, cognitive awareness, automation, and flexible levels of distribution. These aspects facilitate customization to suit different deployment objectives and business models, while advancing a personalized experiential service quality.

2 REFERENCES

[1] NGMN, “5G White Paper,” v1.0”, February 2015.
[2] NGMN Description of Network Slicing Concept v1.0.8, Sep. 2016.
[3] NGMN, “5G End-to-End Architecture Framework,” v3.0.8, September 2019
[4] 3GPP, “System architecture for the 5G System (5GS),” TS23.501 v16.4.0, March 2020
[5] 3GPP, “NR and NG-RAN Overall Description; Stage 2,” TS38.300 v16.2.0, July 2020
[6] ETSI, “ White Paper No.16: GANA (Generic Autonomic Networking Architecture) (http://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp16_gana_Ed1_20161011.pdf)
[7] ETSI, “Network Functions Virtualization (NFV); Infrastructure Overview, GS NFV-INF 001”, V1.1.1, January 2015.
[8] ETSI, “Network Functions Virtualisation (NFV); Ecosystem; Report on SDN Usage in NFV Architectural Framework”, GS NFV-EVE 005, V1.1.1, December 2015
[9] ETSI, “Autonomic network engineering for the self-managing Future Internet (AFI); Generic Autonomic Network Architecture”, TS 103 195-2: (Link:http://www.etsi.org/deliver/etsi_ts/103100_103199/10319502/01.01.01_60/ts_10319502v010101p.pdf)
[10] IETF, “YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF),” RFC 6020, May 2017
[11] The Apache Software Foundation. Apache Kafka. (Link: https://kafka.apache.org)
[12] ETSI, “Open Source MANO”, OSM whitepaper, OSM Release FIVE Technical Overview, January 2019
[13] Xie, M., Zhangy, Q., Gonzalez, A.J., Grensund, P., Palacharlay, P., Ikeuchy, T., Telenor Research, Telenor, Norway, Fujitsu Labs of America, USA. “ Service Assurance in 5G Networks: A Study of Joint Monitoring and Analytics “, 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC): Track 4: Services, Applications, and Business
[14] ETSI, “Core Network and Interoperability Testing (INT); Approaches for Testing Adaptive Networks”, EG 203 341 V1.1.1 October 2016.
[15] GSMA, “Generic Network Slice Template”, Version 3.0, May 2020
[16] NGMN, “A Network Data Layer concept for the Telco industry”, V1.0, August 2018

[17] Cohen, R., Barabash, K., Rochwerger, B., Schour, L., Crisan, D., Birke, R., Minkenberg, C., Gusat, M., Recio, R., Jain, V., " An Intent-based Approach for Network Virtualization", IEEE International Symposium on Integrated Network Management, 2013
[18] van Lingen, F. et al. "The Unavoidable Convergence of NFV, 5G, and Fog: A Model-Driven Approach to Bridge Cloud and Edge." IEEE Communications Magazine 55, no. 8 (2017): 28-35.
[19] Frisiani, G., Jubas, J., Lajous, T., and Nattermann, P., "A future for mobile operators: The keys to successful reinvention", Mckinsey and Company(Link: https://www.mckinsey.com/industries/telecommunications/our-insights/a-future-for-mobile-operators-the-keys-to-successful-reinvention)
[20] NGMN, "NGMN White Paper 2", v1.0, July 2020
[21] ETSI, "Programmable Traffic Monitoring Fabrics that enable On-Demand Monitoring and Feeding of Knowledge into the ETSI GANA Knowledge Plane for Autonomic Service Assurance of 5G Network Slices; and Orchestrated Service Monitoring in NFV/Clouds", PoC (Proof of Concept): ETSI TC INT/ AFI WG, January 2019, (Link: ETSI GANA Model in 5G Network Slicing PoC White Paper #3)
[22] ETSI, " Generic Framework for Multi-Domain Federated ETSI GANA Knowledge Planes (KPs) for End-to-End Autonomic (Closed-Loop) Security Management & Control for 5G Slices, Networks/Services", PoC (Proof of Concept): ETSI TC INT/ AFI WG, May 2020, (Link: https://www.etsi.org/newsroom/blogs/blog-int-Core-Network-and-Interoperability-Testing)
[23] 3GPP, "TR 28.801: Study on Management and Orchestration of Network Slicing for Next Generation Network (Release 15) v1.2.0," 2017.
[24] Letaief, K.B., Chen, W., Zhang, J., Zhang, Y.A., "The Roadmap to 6G: AI Empowered Wireless Networks", IEEE Communications Magazine, August 2019
[25] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, A., Kiddon, C., Konecny, J., Mazzocchi, S., McMahan, H.B., Overveldt, T.V., Petrou, D., Ramage, D., Roselander, J., "Towards Federated Learning at Scale: System Design" Google Inc., Mar 2019
[26] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A.,McMahan, H. B., Patel, S., Ramage, D., Segal, A., Seth, K. Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175–1191. ACM, 2017
[27] Wu, J., Guo, S., Huang, H., Liy, W., Xiang, Y., " Information and Communications Technologies for Sustainable Development Goals: State-of-the-Art, Needs and Perspectives", IEEE, Feb. 2018
[28] UNSDSN. "United Nations Sustainable Solutions Development Network", (Link: https://www.unsdsn.org)
[29] Fettweis, G., "The Tactile Internet: ITU-T Technology Watch Report", August 2014, (Link: https://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000230001PDFE.pdf)
[30] ETSI, "Artificial Intelligence (AI) in Test Systems, Testing AI Models and ETSI GANA Model's Cognitive Decision Elements (DEs)", TC INT/ AFI WG, March 2020, (Link: AI in Test Systems, Testing AI Models and ETSI GANA Model's Cognitive Decision Elements (DEs)) (Link: https://www.etsi.org/newsroom/blogs/blogger/ulrich)
[31] ITU-T, "Distributed ledger technologies: Use cases", Technical paper, HSTP.DLT-UC October 2019
[32] Conti, M., Kumar E. S., Lal, C., Ruj, S., "A survey on security and privacy issues of Bitcoin," IEEE Communications Surveys and Tutorials, 2018
[33] Eyal, I. and Sirer, E.G. (2014), "Majority is not enough: Bitcoin mining is vulnerable", Proceedings of International Conference on Financial Cryptography and Data Security, Berlin, Heidelberg, pp.436–454.

[34] Christidis, K., Devetsikiotis, M., “Blockchains and smart contracts for the internet of things,” IEEE Access, vol. 4, pp. 2292–2303, 2016.
[35] H. Liu, Z. Chen, L. Qian, “The Three Primary Colors of Mobile Systems,” IEEE Comm. Mag., vol.54, no.9, pp.15-21, Sep. 2016.
[36] TMForum, “Blockchain-based Telecom Infrastructure Marketplace”, https://www.tmforum.org/blockchain-based-telecom-infrastructure-marketplace/
[37] TMForum, https://www.tmforum.org/vertical-industry-telcos-federated-dlt-based-marketplace/
[38] McKinsey & Company, “The road to 5G: The inevitable growth of infrastructure cost”, Technology, Media, and Telecommunications, February 2018
[39] Dao, N., Lee, Y., Cho, S., Kim, E., Chung, K., Keum, C., “Multi-tier Multi-access Edge Computing: The Role for the Fourth Industrial Revolution”, 2017 International Conference on Information and Communication Technology Convergence (ICTC)
[40] Tian, F., Zhang, P., Yan, Z., “A Survey on C-RAN Security”, IEEE Access. August 2017
[41] Larsen, L.M.P., Checko, A., and Christiansen, H.L., “A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks,” IEEE Communications, Surveys & Tutorials, Vol: 21, Issue: 1, 2019
[42] 3GPP, “Study on new radio access technology: Radio access architecture and interfaces,” TR 38.801, April 2017, version 14.0.0.
[43] NGMN, “A Network Data Layer Concept for the Telco Industry,” v1.0, August 2018
[44] O-RAN Alliance, https://www.o-ran.org/
[45] “Deutsche Telekom, SK Telecom use network slicing to trial 5G across continents”, Mar. 2018, https://mobileeurope.co.uk/press-wire/deutsche-telekom-sk-telecom-use-network-slicing-to-trial-5g-across-continents
[46] ETSI, “Autonomic network engineering for the self-managing Future Internet (AFI)”, TS 103 195-2 V1.1.1, May 2018
[47] 3GPP, “Architecture enhancements for 5G System (5GS) to support network data analytics services”, TS23.288 V16.3.0. March 2020
[48] P. Jamshidi et al., “Microservices: The Journey So Far and Challenges Ahead”, IEEE Software, vol. 35, no. 3, May/June 2018, pp. 24–35.
[49] Rajan, D., “Common Platform Architecture for Network Function Virtualization Deployments”, 2016 4th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering
[50] FCC, “Promoting Investment in the 3550-3700 MHz Band,” GN Docket No. 17-258 October 2018).
[51] 3GPP, “Study on application architecture for enabling Edge”, TR 23.758, v17.0.0, December 2019
[52] ETSI, “Cloud RAN and MEC: A Perfect Pairing”, Whitepaper 23, February 2018.
[53] Yosra, S.B., Allio, S., Jacques, J., “Anomaly Prevision in Radio Access Networks Using Functional Data Analysis”, IEEE Global Communications Conference (GLOBECOM), 2017
[54] 3GPP, “Study on Security Impacts of Virtualisation”, TR 33.848, v0.5.0, November 2019
[55] 3GPP, “Security Assurance Methodology (SECAM); and Security Assurance Specification (SCAS); for 3GPP virtualized network products”, TR 33.818, v0.7.0, May 2020
[56] 3GPP, “Study on authentication enhancements in 5G System, TR 33.846, v.0.7.0, August 2020
[57] 3GPP, “NR and NG-RAN Overall Description”, TS 38.300, v16.3.0, September 2020
[58] 3GPP, “Security architecture and procedures for 5G system”, TS 33.501, v16.4.0, September 2020
[59] 3GPP, “NG-RAN: Architecture Description”, TS 38.401, v16.3.0, September 2020
[60] 3GPP, “Study on Security Aspects of Enhancement of Support for Edge Computing in 5GC”, TR 33.839, v0.1.0, August 2020

[61] 3GPP, “Generic Bootstrapping Architecture (GBA)”, TS 33.220, v16.2.0, September 2020
[62] 3GPP, “Battery Efficient Security for very low Throughput Machine Type Communication (MTC) devices (BEST)”, TS 33.163, v16.2.0, September 2020
[63] IETF, “A Privacy Mechanism for the Session Initiation Protocol (SIP)”, November 2002
[64] IETF, “Private Extensions to the Session Initiation Protocol (SIP) for Asserted Identity within Trusted Networks”, November 2002
[65] 3GPP, “Study on 5G Security Enhancement against False Base Stations”, TR 33.809, v0.10.0, August 2020
[66] 3GPP, “Study on evolution of Cellular Internet of Things (CIoT) security for the 5G System”, v16.1.0, August 2020
[67] 3GPP, “Study on security aspects of 3GPP support for advanced Vehicle-to-Everything (V2X) services”, v16.1.0, September 2020

3 DEFINITIONS

AF Autonomic Function, which is a type of function that does not require configuration, (except for being subject to input governance policies and the setting of operational mode (Open loop or Closed loop) and is able to derive all the necessary information, through self-knowledge, discovery, or policies.

AI and ML Data Model A model representing mathematical algorithms that learn using data and input consisting of human expertise to generate an effective and optimized decision, in the presence of evolving complexity, when the model is provided with actual information of a corresponding nature for which the model was designed

E2E End-to-End, which refers to communications between two endpoint devices or user equipment, across any arrangement of intervening administrative domains

DLT Distributed Ledger Technology. This is a distributed database that leverages the blockchain framework, for storage, access, and adding data records securely, by authorized entities.

Haptic Sense Haptic sense is perception characterised by touch. This type of perception is associated with tactile sense (derived from the Latin: *Tangere* - to touch), and kinaesthetic sense (derived from the Greek: *Kinesis* – movement, and *Aesthesis* – perception), for example body movement.

Microservice A small self-contained service, which is a collection of smaller loosely-coupled components

Network Function (NF) Processing function in a network. This includes a variety of control plane, user plane, and service functions that span the layers of the protocol stack. (e.g. radio network functions, physical layer functions, Internet Protocol (IP) routing functions, applications etc.) [2]

Network Service Provider (NSP) Entity that provides network access service and owns related resources and functions (e.g. virtualised or physical) for providing network access. The resources and functions include spectrum, mobility, and access management across heterogeneous and/or composite access networks, network management and orchestration, and network elements [2]

Network Slice Blueprint (NSB) A complete description of the structure, configuration, and the plans/work flow for how to instantiate and control the Network Slice Instance during its life cycle. A Network Slice Blueprint enables the instantiation of a Network Slice, which provides certain network characteristics (e.g. ultra-low latency, ultra-reliability, value-added services for enterprises, etc.). A Network Slice Blueprint refers to required physical and logical resources and/or to Sub-network Blueprint(s)

Network Slice Instance (NSI) A set of run-time network functions, along with physical and logical resources to run these network functions, forming a complete instantiated logical network to meet certain network characteristics required by the Service Instance(s). A network slice instance may be fully or partly, logically and/or physically, isolated from another network slice instance [2]

Service Instance (SI) An instance is a run-time construct of an end-user service or a business service that is realised within or by a Network Slice [2]

Service Provider (SP) Entity that provides an application layer service. The entity may be a third-party, or an NSP.

Tactile Internet (TI) Tactile Internet consists of a variety of assorted and hybrid cyber-physical interfaces, such as haptic (sense of touch), proprioceptive (sense of perception characterized by a combination of body position and movement), visual (ocular sense), audio (hearing sense) etc., which require ultra-low latencies and high reliability human and machine interfaces

X-Haul A common flexible transport solution for future 5G access networks, which aims to integrate fronthaul and backhaul networks with all their wired and wireless technologies in a common packet-based transport network under SDN (Software Defined Network) based and Network Functions Virtualisation (NFV) enabled common control.

4 AUTONOMIC NETWORKING

4.1 General

A generic autonomic networking architectural model was introduced in [3] leveraging the concepts described in [6][9] which utilizes two hierarchically oriented, nested control-loops, characterized by time-scale and objectives for autonomic management and control. This model is further elaborated in terms of a variety of usage scenarios that leverage autonomic networking. The principles embodied in autonomic networking combine cognitive awareness with self-adaptive functions, over a system-wide and virtualized environment, to enable a variety of flexible implementation and deployment strategies. An extensible 5G context for autonomic networking includes the various aspects of network slicing from an end-to-end system perspective, as depicted in Fig. 1.

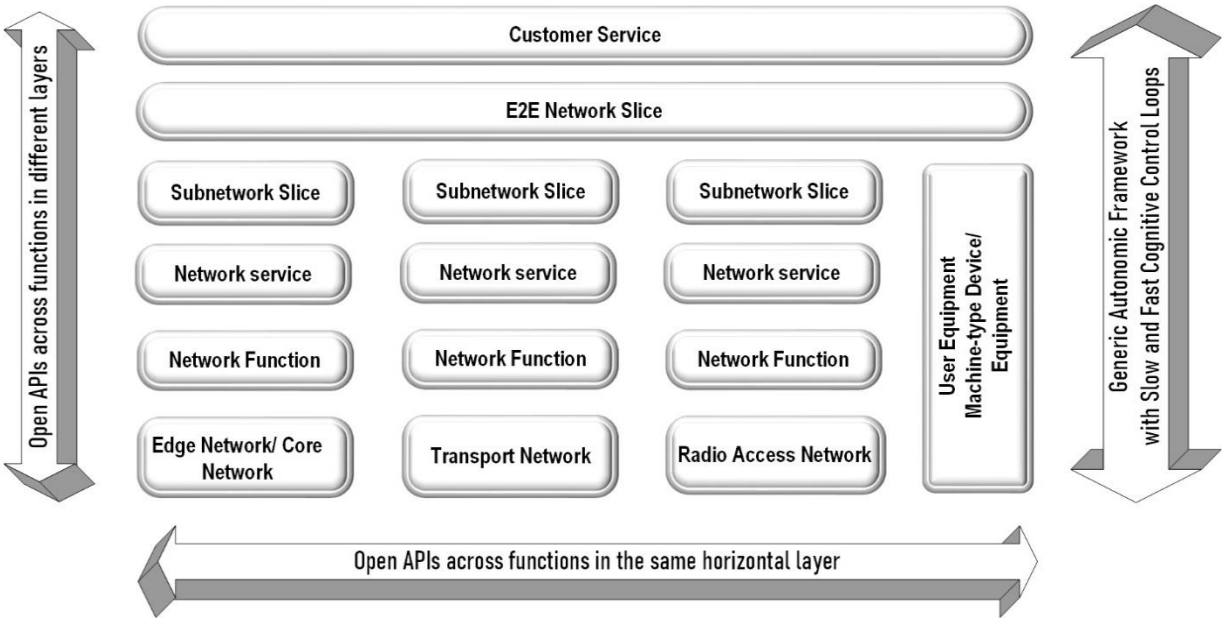


Fig. 1 : Extensible 5G context for an autonomic networking framework

The flexible and diverse service paradigm envisioned in 5G requires a correspondingly supportive network advancement, in terms of efficiency, resource optimization, performance, quality, reliability, migration choices, business models and availability. This requires a framework for the management of the proportionately increasing levels of system complexity. Complexity in such an environment is a result of the characteristics of service flexibility and diversity, together with the heterogeneity of radio access networks, plurality of deployment arrangements, virtualization, inter-domain network cooperation, common marketplace of services etc. The interdependent and dynamic nature of these characteristics further compound the nature and level of complexity, which requires to be effectively and efficiently managed through automation on a system-wide basis.

Autonomic networking provides a scalable, customizable, and self-organizing model for system-wide automation, where virtualization and programmability are adopted in an end-to-end manner. Self-organization is an integral aspect of autonomic networking for automation that dynamically learns and adapts to changing conditions in the system and its environment, while satisfying service KPIs and the quality of service experience.

The virtualization [7] of system resources, consists of network, computing, and storage, together with programmability [8], through the adoption of software defined networking that embodies a separation of the

control plane and user plane. The composability of virtualized functions facilitates a dynamic, on-demand, configuration, and instantiation of logical networks through the enabling construct of network slicing, which is a pivotal ingredient of system-wide virtualization. The separation of the control plane and the user plane, the shared network data layer, and the use of stateless functions in the network, together with heterogeneity and dis-aggregation of the radio access network, facilitate customizable levels of flexibility and granularity in a virtualized context.

With these inter-dependent, system-wide requirements, automation is a significant capability for adequately supporting the system behaviour in a predictable and sustainable manner, in the presence of a dynamic and changing environment of services and system conditions. Network services, subnetwork services, network slice services and customer services that operate in a virtualized end-to-end system are facilitated through appropriate levels of CI (Continuous Integration) and CD (Continuous Delivery) of services over cloud-native constructs, where services are composable through malleable combinations of smaller components, known as microservices. This model for service creation and rendering is amenable for automation, through autonomic networking that leverages analytics and machine-learning. Autonomic networking consists of self-CHOP (Configuration, Healing, Optimization, and Protection) capabilities. These capabilities provide an optimization of resource utilization, performance, and quality of service experience. Such optimization is pivotal to effectively manage the scale and complexity of an expansive 5G ecosystem of heterogeneous connectivity (fixed and mobile access), plurality of user equipment, and the diverse demands of a wide-variety of services.

Autonomic attributes are descriptive of any autonomous system, where cognitive awareness is realized through various modalities of AI (Artificial Intelligence) and ML (Machine Learning). Cognitive awareness within virtualized functions is amenable to programmability and ease of deployment, through the architectural tenets of SDN (Software Defined Networks), while allowing flexible arrangements of distribution and decentralization to support edge computing. Flexibility and granularity of deployment arrangements, service creation and rendering, is accomplished through the composability any given service in terms of microservices, within a service-based framework. The programmable context of a service-based framework serves as a cloud-native environment for the adoption of autonomic behaviors, where the flexibility, granularity and agility of functions is rendered through the composability of constituent services. Open and Representational State Transfer oriented Application Programming Interfaces (REST-ful APIs) provide for interactions among cooperative Virtual Network Functions (VNFs), through the service exposure layer, both vertically across the layers, and horizontally within any given layer, of the SBA (Service Based Architecture) [4], which enables a service based framework. The logical nature of the service based framework model, is relevant for the Core Network (CN) and the edge network, which embodies the characteristics of Multi-access Edge Computing (MEC).

4.2 High level architecture

The 5G and beyond ecosystem is characterized by a prolific evolution and expansion of the service paradigm, over a distributed heterogeneous access with variable coverage area footprints, and multiple access technologies coexisting and cooperating with the advancing capabilities of New Radio (NR) access [5]. Distribution and decentralization to optimize the system performance and the user experience, are enabled through a closed-loop feedback control system of cognitive capabilities with system wide scope of awareness, broadly referred to as the Knowledge Plane (KP). Flexible deployment arrangements along a forward-looking direction of localization, resulting from decentralization and distribution, are facilitated by MEC, network convergence, and integrated access and backhaul.

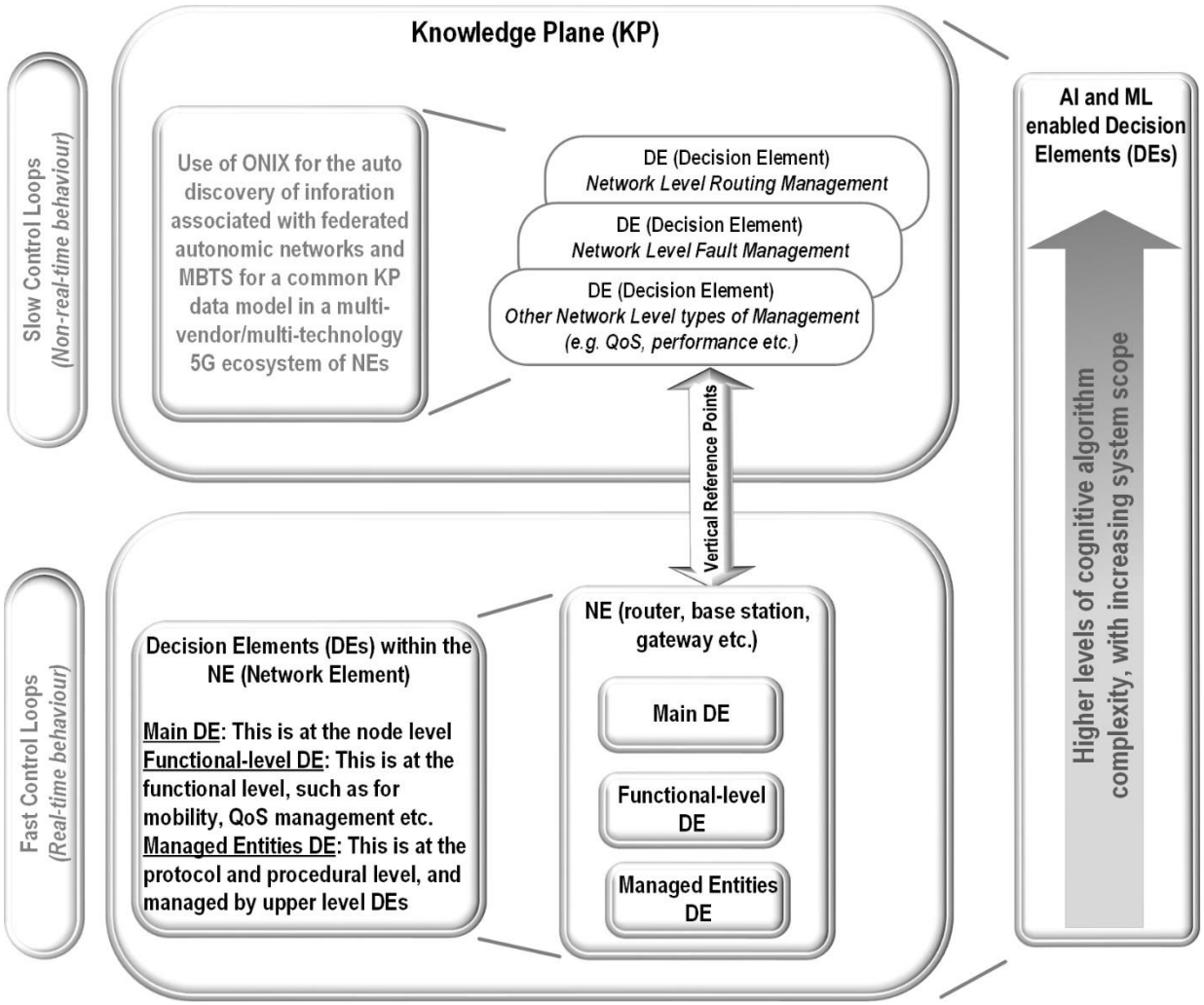


Fig. 2 : High-level architectural model of AI and ML enabled KP

The high-level architectural model of an AI and ML enabled KP constitutes the fabric of autonomic networking depicted in Fig. 2, which provides a foundation for automating the operation and performance optimization of an end-to-end system consisting of a core network, edge network, access network, and user equipment. Virtualized functions, supported by generic computing, storage, and networking resources, leverage the constructs of SDN, across a multi-vendor landscape. Autonomic networking operates over a context of network slicing, which supports the requirements of multiple tenants, including Verticals, different domains etc., using specific NSIs (Network Slice Instances) over shared computing, storage, and network resources in the system.

The layered architectural model of the end-to-end system facilitates flexibility in terms of business models, deployment arrangements, across a single or multiple NSP and/or SP domains, allowing a variety of federated partnerships. Consistent system behaviours and optimized service experience, are among the benefits of federated partnerships, across topologically disparate NSP and/or SP serving areas.
The different layers of the architectural model performs the functions that characterize the essential capabilities in a virtualized end-to-end system, where autonomic networking serves as a cross-layer sub-system for effectively managing complexity, while offering a requisite network slice for any tenant of the network slice in a multi-tenant system. The virtualized nature of an end-to-end system is realized as virtual machines that virtualize the hardware, executing on different instances of the operating system, or as

containers. In the case of containers, the operating system is virtualized to enable a variety of workloads or processing demands that execute over a single operating system instance. Relative to a virtual machine, a container is light weight and portable since the operating system is common for the different containers. These realizations allow for a multi-vendor participation, and deployment arrangements that are distributed.

The different planes in the service based framework that cooperate and collaborate with one another provide the operational context for the KP in autonomic networking as shown in Fig. 3.

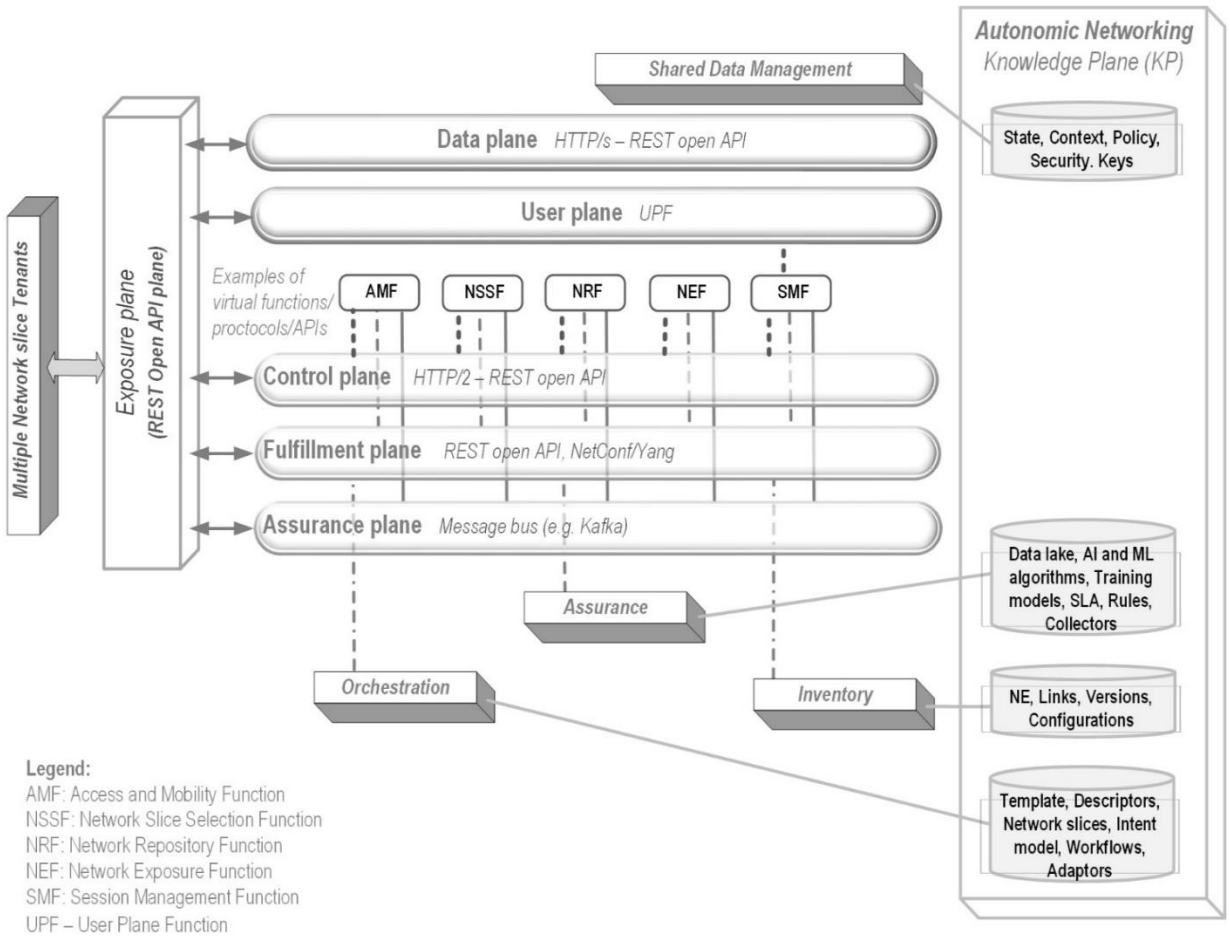


Fig. 3 : Illustration of a service based framework with autonomic networking

Autonomic networking containing the KP, utilizes a distributed arrangement of fast and slow feedback loops shown in Fig. 2, in a system-wide manner, across the layered architectural model shown in Fig. 3. for applicability to a single NSP domain or across multiple NSP domains. For example, autonomic networking manages complexity and automates interactions among multiple stakeholders (e.g. NSP, SP), such as in a common marketplace usage scenario described in section 9.5.2, or in customized resource sharing deployment arrangements, with relevant service level agreements, associated with a diversity of tenants, depicted in Fig. 3.

The different planes, and the management aspects consisting of orchestration, assurance, inventory, together with the KP in autonomic networking, as depicted in Fig. 3, which are virtualized, Virtual Machine (VM) based, or containerized, multi-vendor-oriented, and distributed, are layered from an end-to-end network perspective. The end-to-end network may span heterogeneous access with fixed, broadband, license-exempt

access, transport network with SDN, edge and core network, non-terrestrial access, non-public network, IMS network etc.

The logical layers of a service based framework depicted in Fig. 3 support the following system and sub-system layers, as shown in Fig. 4, and aligned with Fig. 1:

- Customer service layer
- End-to-End network slice layer
- Subnetwork slice layer
- Network service layer
- Network function layer
- Infrastructure layer consisting of the end-to-end network

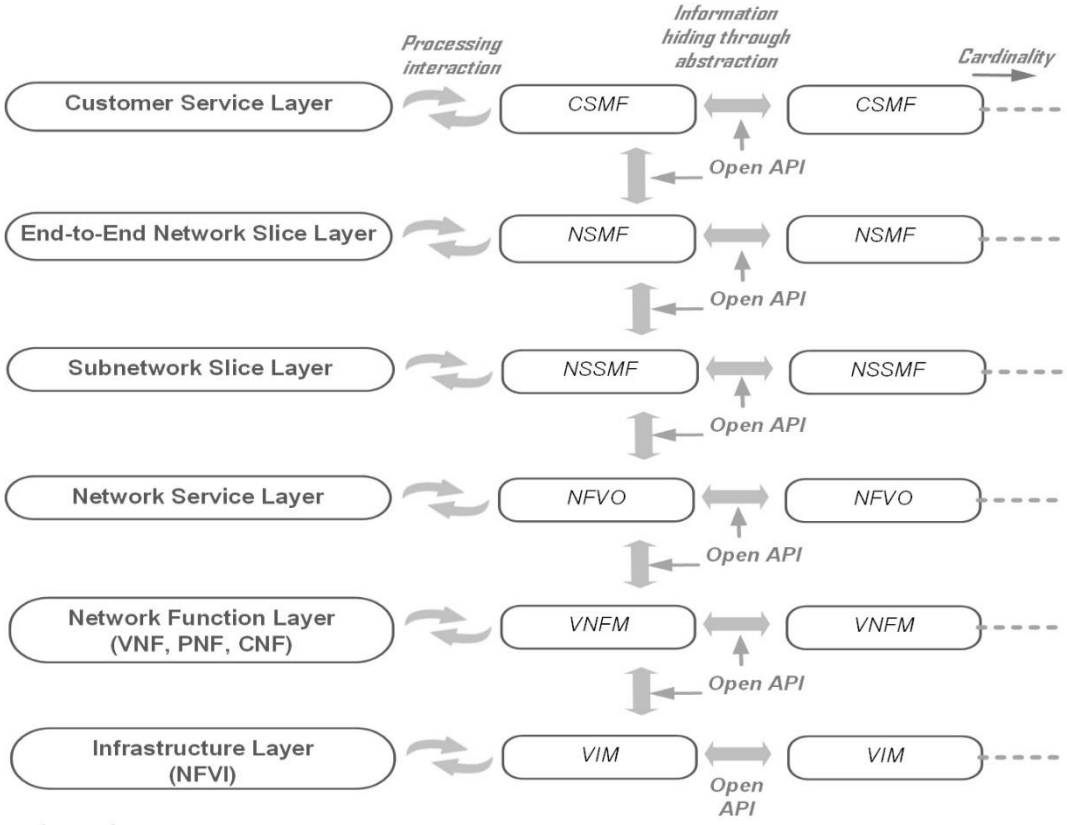


Fig. 4 : Logical layers from an end-to-end network perspective

The service based characteristics of the end-to-end system is underscored by the cloud-native characteristics inherent in the flexibility afforded by microservices, which facilitates a composition of higher-order services, using smaller constituent components. With the use of stateless functions, the design and implementation of virtual network functions is simplified, while enhancing scalability, since then it is unencumbered by storage considerations.

Each layer is managed and orchestrated with different phases of design, instantiation and configuration, activation, monitoring and decommissioning and some level of automation, closed loop feedback control, and with open APIs: the vertical APIs between the layers, and the horizontal APIs at the same layer (e.g. across different domains).

4.2.1 User Plane

This plane is associated with the UPF, together with the storage and retrieval of data in a shared data environment, further elaborated in section 5..

4.2.2 Control Plane

This plane, via the SMF, is associated with the control of the user plane as well with the authentication, security, policy, charging network functions. It supports the create-delete-subscribe-notify operation on information, with respect to the network functions in the service-based framework.

4.2.3 Data Plane

This plane is associated with the underlying fabric data, which serves as the glue (e.g. HTTP/s, REST-ful API) for information exchange across the end-to-end system.

4.2.4 Fulfillment Plane

This plane is associated with the provisioning and configuration of the system, while it also provides the information pertaining to the configuration.

4.2.5 Assurance Plane

This plane is associated with the issuance of information pertaining to events, logs, alarms, performance metrics, as well as to detect incidents, analyze, and operationalize to guarantee the quality of service.

4.2.6 Exposure Plane

This plane is associated with the exposure of data, control plane functions, provisioning or ordering functions, management functions or events, and to provide command and control APIs for different tenants.

The KP interacts with these different planes to store and retrieve data, such as templates, descriptors, service description, policies, information models, state of network functions etc. Each network function, whether it is a core network function or a management network function, interacts with these planes, as needed, over open and interoperable interfaces. The KP requires to be multi-tenant, secure and persistent. From a MEC perspective, the KP instances are distributed, as needed, for providing ultra-low-latency and high reliability services over a heterogeneous edge network, with appropriate information storage capacity, for an advanced service quality and experience.

At each layer, the KP enabled autonomic networking provides system wide capabilities to manage, configure, activate, orchestrate, instantiate, monitor, and decommission the resources associated with the establishment of an end-to-end network slice. This allows for an automation and optimization to adequately support the service demands (KPIs), of a given machine-type equipment or end-user equipment. In other words, autonomic networking provides the capabilities for automating the E2E system behaviour, while allowing flexible system deployment scenarios to suit diverse business models. This process leverages the use of open REST-ful APIs, in the vertical direction across the layers, and in the horizontal direction within a given vertical layer in terms of processing, communication, and storage resources (e.g. core network, edge network, radio network, and device), as well as across disparate administrative domains. The complementary attributes of the composable layers and open APIs are required to be specified as interoperable capabilities, for a realization through the use of programmable methods, such as SDN oriented core and MEC networks.

4.2.7 Service Orchestration

The delivery of an optimal service quality and experience is the objective of service orchestration, where the different constituents of an end-to-end system cooperate in terms of a dynamic and optimized workflow. Service orchestration translates “what is required” into “how the requirements are satisfied”, and the realization of a corresponding configuration of resources for a network slice, based on a given data model abstraction (e.g. YANG) [10] associated with a tenant, virtual function, analytics etc. The enforcement of a configuration occurs through the use of NETCONF [10] interfaces exposed within a cloud-native core, edge, transport, or radio access network. The use of NETCONF as part of service management, has been widely adopted as a standardized protocol in the industry.

Service orchestration, augmented by autonomic networking capabilities, provides for high levels of automation by coordinating diverse interactions associated with service workflows among network entities, consisting of VNFs and Physical Network Functions (PNFs) in the end-to-end system, where at the 5G network layer the various lifecycle management aspects, such as inventory, modeling, on-boarding, monitoring etc. of the constituent network entities, are supported. Dis-aggregated network entities in the RAN include hybrid combinations of VNF and PNF (e.g. antennas, which have PNF resources pertaining to physical geometry etc.)

Directions for SDN controllers, which facilitate the programmability of the service based framework for a diversity of usage scenarios, are provided by a service orchestrator. The different layers of the service based framework, which consists of an implicit hierarchy with a distributed and decentralized architectural arrangement, include virtual and physical resources in the core network, edge network, radio network and the transport network. In this architectural arrangement, unique realizations of an end-to-end network slice is established through a selective utilization of virtual and physical resources to meet the requirements and KPIs associated with a given service tenant. This approach enables a logical sharing of the network infrastructure as depicted in Fig. 3, via end-to-end network slicing, where each network slice is configured and instantiated to support the requirements of a service tenant.

The service orchestrator may consist of sub-system level orchestrators that cooperate to aggregate the required virtual and physical resources, over a single API. The SDN controller at any given layer exposes the abstract presentations of the virtual and physical resources utilized by the service orchestrator. The service orchestrator decomposes requests from a given service tenant into the required virtual and physical resources and conveys the corresponding request to the appropriate SDN controller for acquiring the required virtual and physical resources. The service orchestrator may also be capable of cooperating with other service orchestrators, across multiple domains, where the instantiation of an end-to-end network slice would require the fulfillment of the associated inter-domain Service Level Agreements (SLAs), which are to be mapped into a corresponding allocation of virtual and physical resources, service quality requirements, as well as the lifecycle management of service tenants.

A service orchestrator having a multi-domain, multi-layer scope, cooperates with the orchestrators associated with the different sub-systems consisting of the core network, edge network, and the radio network, for a cohesive workflow, in a multi-vendor ecosystem, for the realization of an end-to-end network slice. This type of architectural arrangement allows for a flexible scaling up or down of virtual and physical resources to suit the performance KPIs associated with a service tenant, while optimizing the efficiency and utilization of resources. For example, a multi-domain orchestrator interacts cooperatively with a MEC orchestrator at the sub-system level, in the presence of dynamic load conditions (e.g. network, processing, storage etc.), to adjust to the system performance to satisfy the diverse demands of a service tenant (e.g. tactile internet, industrial automation etc.).

Microservice based orchestration allows for a distribution of the orchestration process across distributed resources, which allows for a flexible partitioning of the orchestration process, with loose coupling and high

cohesion to suit a given service deployment choice. The microservice based orchestration is based on a publish-subscribe scheme among the distributed components of the orchestration process, using high scalable event handling subsystems (e.g. Kafka [11] [12]).

Some of the significant aspects of service orchestration, augmented by autonomic networking include the following:

- Management of an intent based framework, which emphasizes a desired state, instead of specific actions to get to the desired state, to satisfy the diverse demands of service tenants in terms of SLAs, through a requisite mapping of virtual and physical resources to an end-to-end network slice.
- Onboarding of software templates and packages with CI-CD, together with an analysis of these software templates and packages, which are stored in common repositories.
- Construction of service logic to perform feasibility checks and to build the different configuration and instantiation requests for the underlying sub-systems, via open APIs or adaptors that allow a mapping of the virtual and physical resources to specific underlying technology implementation.
- Updating of the service catalog with any new services that are available for service tenants, and the associated end-to-end network slice instance.
- Ordering process that allows a service tenant to order a new service instance from the service catalog, and the corresponding activation of the order request by triggering the instantiation of the supporting end-to-end network slice instance.
- Instantiation of an end-to-end network slice instance with the provisioning of the required configuration and activation of virtual and physical resources in compliances with the policy, security, and KPI profiles for the service tenant (i.e. customer)
- Closed loop activation and instantiation process with service assurance in the loop to ensure anomaly detection, mitigation, and resolution, with respect to constituent sub-network slices and associated domains that are engaged in the composition of an end-to-end network slice instance.
- Monitoring, collection, and analysis of information associated with the optimization of operation in the service environment.

4.2.8 Service Assurance

Service assurance supports the orchestration and provisioning of network slices to suit the KPI profiles for a service tenant (i.e. customer), in a service-based framework. The service assurance architecture consists of distributed and modular monitoring and analytics, for an end-to-end network slice, across the core network, edge network, transport network, and the radio network, to support a diverse variety of KPI profiles, adequately and flexibly, in a multi-tenant service environment. These capabilities of service assurance are applicable in intra-domain and inter-domain arrangements.

Closed-loop feedback operation across service orchestration, provisioning, and service assurance is depicted in Fig. 5.

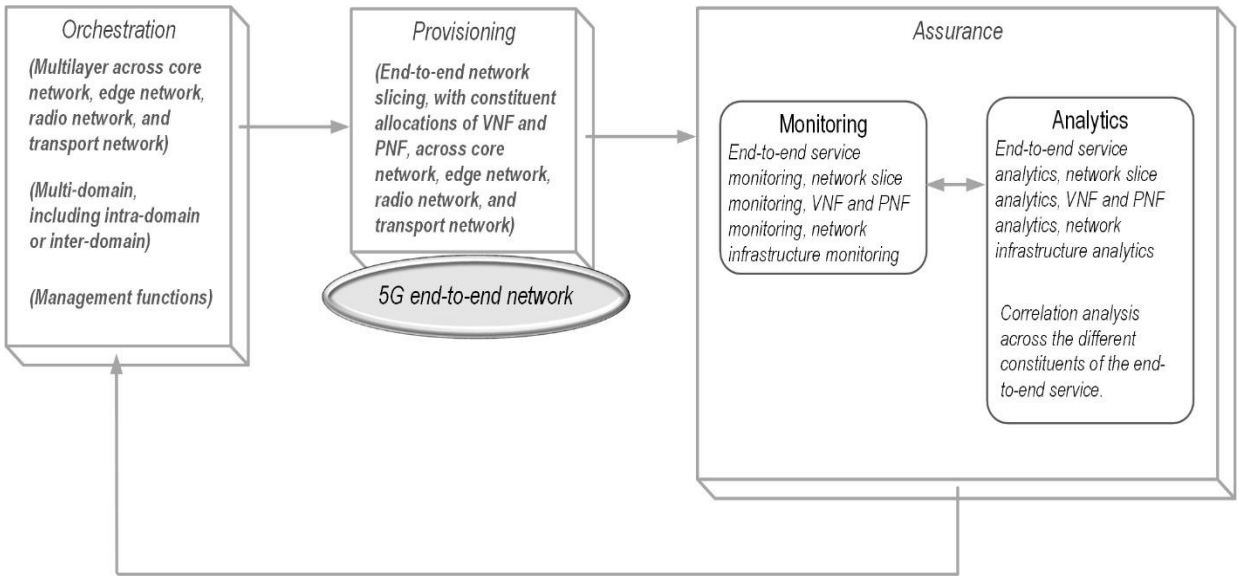


Fig. 5 : Closed-loop collaboration between service orchestration and service assurance

The closed-loop feedback between service orchestration and service assurance, facilitates the optimization of resources for the provisioning of an end-to-end network slice for a customizable delivery of service requested by a tenant. In this context monitoring and analytics cooperate, in accordance with policies and service requirements (e.g. SLAs, KPIs etc.) for a dynamic allocation of the requisite resources that comprise a given end-to-end network slice.

A modular and distributed approach for service assurance, accommodates autonomic networking principles for self-organizing behaviours that allow an automatic and flexible provisioning of resources, via service orchestration. From a tenant (i.e. customer-centric) perspective the monitoring of the Quality of Experience (QoE), provides the appropriate inference of the quality service experience as perceived by an end-user, based on the quality of the underlying end-to-end network slice. The service assurance for an end-to-end network slice monitors and analyses the resources associated with each segment of the end-to-end network (e.g. core network, edge network, transport network, and radio network). For an effective assurance of an end-to-end network slice service, a coordination of monitoring and analysis is pivotal for cross layer process transactions, such as for example in the case of containerized microservices, which would require the use of an Open API framework to facilitate a vendor neutral open interface to different implementations, in a cloud-native environment. The use of a closed-loop feedback model depicted in Fig. 5 is of paramount significance for a dynamic and intelligent service assurance, for advanced end-user services, associated with an emerging ecosystem of Vertical markets.

Within the monitoring aspect of service assurance, the inference of the performance of a given end-to-end network slice is typically derived from an observation of selected traffic flows. For example, some of the traffic flow selection methods include network tomography [13] that leverage Maximum Likelihood Estimator (MLE), and Bayesian inference, where the performance of a given end-to-end network slice is inferred through inferential methods of monitoring. The benefit of a collaborative monitoring and analysis within service assurance is the ability to identify a root cause associated with a detected anomalous behavior of an end-to-end network slice, where the constituent resources (e.g. VNF, PNF, processing, network, spectrum, storage etc.) of the end-to-end network slice are distributed across the different segments of an end-to-end network (e.g. core network, edge network, transport network, radio network).

From an autonomic networking perspective, cooperation and collaboration of monitoring and analysis, with service assurance provides the requisite cognitive intelligence to satisfy the quality of service

requirements the related SLAs, and the system policies, aligned with the demands of a customer. Closed-loop feedback operation, with service orchestration, in conjunction with AI and ML algorithms leverage alarms, notifications, and storage in a common data lake to trigger appropriate corrective actions that promote self-organizing behaviors. Some of the prominent features, and aspects associated with an intelligent service assurance include:

- Analysis of data with the use of AI and ML algorithms, for building smart metrics, pattern recognition, and anomaly detection
- Onboarding and validation of new AI and ML models, algorithms, and rules.
- Update of graph databases, which provide a global view of the structure of an end-to-end network in terms of resources that could be leveraged in the orchestration, provisioning, and instantiation of an end-to-end network slice.
- Dashboard, statistics, and reports
- Fault and SLA management
- API handler for different APIs for onboarding, exposure, or interaction with other entities, including closed-loop feedback with service orchestration

4.3 Closed-loop end-to-end system

Autonomic networking operates through the use of feedback loops, which are overlaid with the AI and ML models that are applied across entities in the service-based framework for imbuing self-organizing capabilities in cooperation with higher levels of abstraction in the KP. The feedback loops are at a building-block level of granularity, where the operation is triggered by an event associated with a given entity, which then initiates an analysis, and subsequently a feedback command to the original entity, via a notification to an orchestrator.

4.3.1 Generic feedback model

A generic feedback model within autonomic networking from an end-to-end perspective follows a closed-loop sequence that starts with events in the network infrastructure entity, analysis in the intelligent assurance entity, which notifies the service orchestrator that appropriately commands the network infrastructure for adaptive and optimized behaviours. A generic feedback model, with participating entities is depicted in Fig. 6.

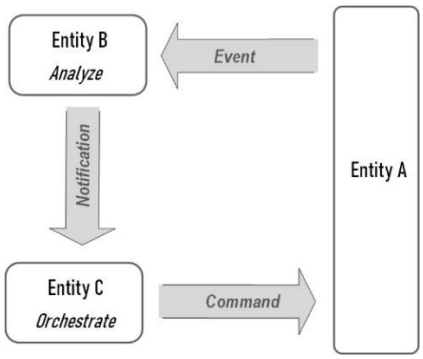


Fig. 6 : Closed-loop feedback model for system entities

For example, the information collected by AI and ML assisted assurance or intelligent assurance, such as logs, alarms, and metrics are analyzed based on AI and ML algorithms for anomaly detection, with appropriate notifications that are send to the service orchestrator. Graph databases are utilized, for modeling rules or algorithms associated with a complex or unstructured environment, such as in a network with dynamic relationships between related data (e.g. heterogeneous radio access etc.).

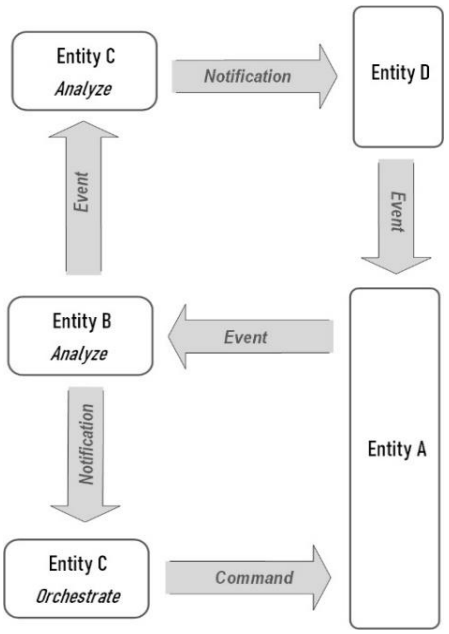


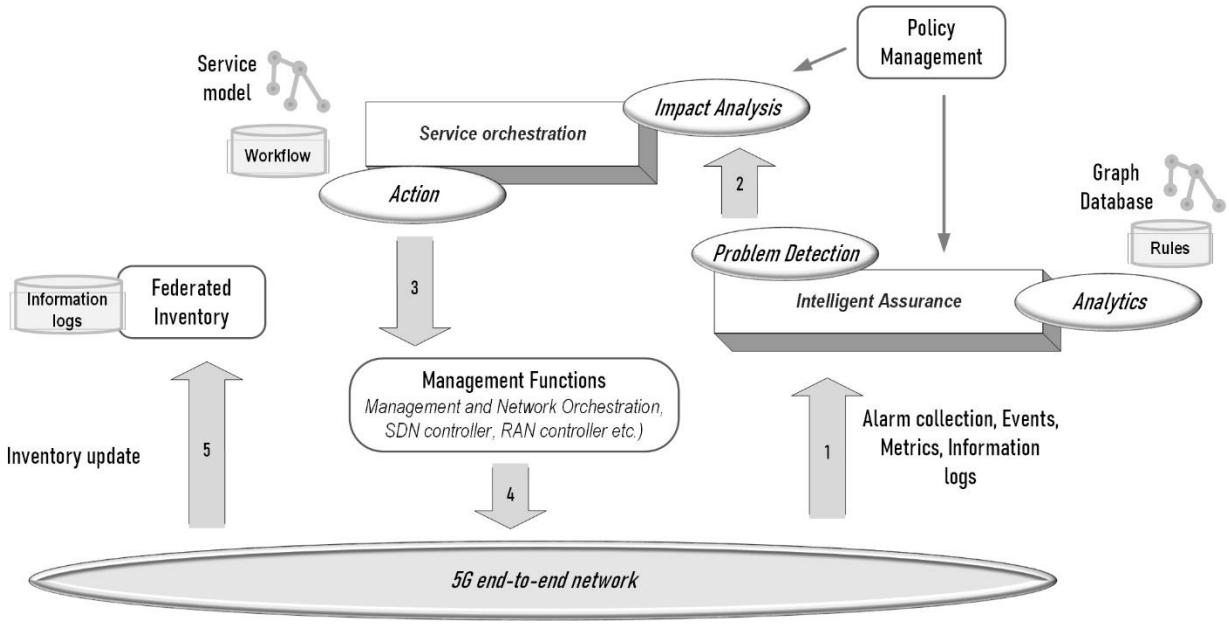
Fig. 7 : Multiple closed loop feedback model

Service orchestration analyzes the impact of the notification on the overall service, and based on the service model, the policies, and the workflows, the orchestrator composes the appropriate action, via commands with relevant system or sub-system scope, for the autonomic networking system, to guide the network infrastructure towards self-organizing and optimized behaviours. After the network infrastructure response is complete, the federated inventory is updated.

A feedback loop associated with a given set of entities, may also interact as needed with another feedback loop to control and guide the behaviour of a group of interacting entities within an end-to-end system, as shown in Fig. 7.

4.3.2 Feedback control for autonomic behaviour

The dynamic nature of feedback loops within autonomic networking is characterized in terms of a variety of factors at the time of network slice instantiation, such as the number of users, the services being rendered, the traffic conditions in the network, where the network infrastructure behaviour is correspondingly affected. The decisions made through cooperative, cognitive feedback loops within autonomic networking, facilitate a correction [14] of anomalies that are detected and an establishment of optimized performance, for a given context, in the networking infrastructure from an end-to-end perspective.



:

Fig. 8 : System context for autonomic networking operation

The system context for autonomic networking operation is shown in 0.

4.4 Usage scenarios

The service orchestration, provisioning, and assurance of an end-to-end network slice with closed-loop, self-organizing behaviours to adapt dynamically to changing network conditions, and end-user service demands, provides the support for multiple Vertical markets. As the end-to-end network evolves in terms of technology as well as flexible deployment arrangements, the objective of a cooperative and collaborative service orchestration, provisioning, and assurance is anticipated to expand progressively from a connectivity-oriented model to support usage scenarios that include virtual functions, processing, networking, storage, and end-user facing services, in a multi-tenant service-based framework.

A few examples of customer usage scenarios that leverage the closed-loop facilitated self-organizing behaviours of service orchestration, provisioning, and assurance include:

- ***Onboarding of a network slice template:*** A standardized network slice template, such as that defined by GSMA NEST [15], and onboarded via an open API on the service provider Network Slice Management Function (NSMF), which provides the management services for one or more Network Slice Instance (NSI), and may consume some management services produced by other functional blocks. The network slice template is then validated and stored in a common database automatically. If an error is identified, a corresponding error message is sent back to the onboarding API.
- ***Service assurance:*** Metrics are collected on the infrastructure, the network functions, the network slices and stored in service assurance to be analyzed. If an abnormal behavior is identified, a message is sent to the problem resolution module that then recommends an action to the service orchestrator, automatically.
- ***V2X service:*** Events are collected from connected cars and sent to an edge application. If a road hazard is detected, a message is broadcast to the surrounding cars to inform them about this road hazard.
- ***Broadcast service:*** A video recording is sent local studio application where it is rendered, and then broadcast accordingly to different types of devices

5 NETWORK DATA LAYER

5.1 General

The management of network data is a critical aspect of the Operational Support System (OSS) for an efficient enablement of cloud-native architectural tenets associated with Network Function Virtualization (NFV). Both integrated and distributed database arrangements are required for a flexible realization of VNFs (Virtual Network Functions). In this context the establishment of a separate NDL (Network Data Layer), managed by the OSS, enables easier, faster, and flexible deployment choices for VNFs, while optimizing operational efficiency, for diverse usage scenarios.

The separation of application logic associated with data processing, from the data storage that may be centralized or distributed, is an intrinsic characteristic of the VNF in the NFV framework. In other words, VNFs can be stateless, delivering only the required service business logic, without managing their own data. This separation enables directions towards an independent scaling of the heterogenous data processing and the data storage requirements, while promoting deployment flexibility, promoting system-wide efficiency, in concert with autonomic networking enabled self-organization, to reduce the Total Cost of Ownership (TCO). The leveraging of autonomic networking within cloud-native architectural tenets, facilitates the management of complexity through system-wide automation and self-organizing attributes.

5.2 Architectural considerations

Autonomic principles embodied within the system-wide scope of autonomic networking utilize AI and ML enabled cognitive capabilities for data processing and analytics for the self-organization and performance optimization of diverse network deployment architectures. In other words, the separation of data processing and data storage, with respect to network data, is augmented through the automation of the OSS enabled by autonomic networking.

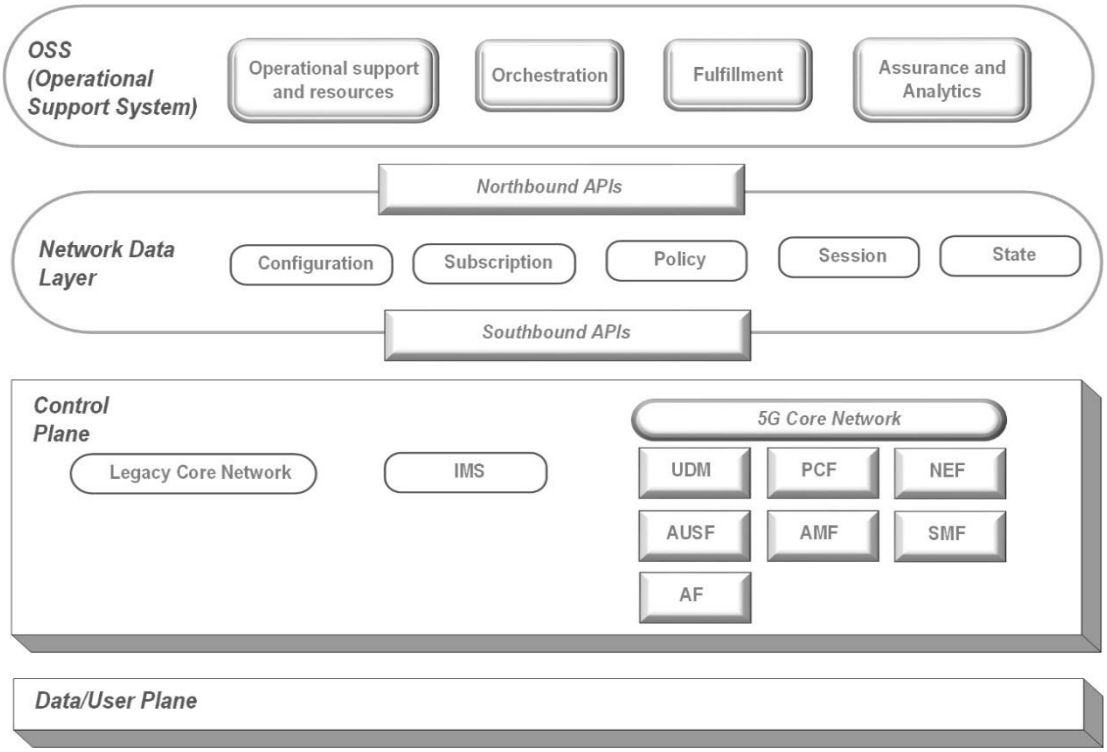


Fig. 9 : NDL architectural context

A contextual representation [43] of the NDL is shown in Fig. 9. The enablement of support for both stateful and stateless services requires a service based framework, built on NFV principles, which not only include VNFs, but also provides a path for transitioning away from proprietary storage hardware, provisioned and configured in silos associated with different applications and service, towards software-defined storage over shared physical storage resources [49]. This approach facilitates the realization of a policy-oriented automation that scales the storage to meet the demands of storage in the NDL.

This intelligent, and virtualized storage capability in concert with embedded cognitive capabilities facilitated in cooperation with autonomic networking supports a context aware storage behaviour to suit the service demands from an end-to-end perspective.

5.3 Stateful and Stateless Considerations

Stateful applications and processes maintain their state within themselves, for the persistence of an associated context. The persistence of context is particularly beneficial for any session-oriented application or process, such as voice, video etc., to preserve the associated session relation information, such as users, location, devices, codecs, duration, resources etc. Consequently, the state information is lost if the underlying process that preserves the state information terminates.

On the other hand, stateless applications and processes are independent of a preservation of their state for consistent and predictable behaviours. A prominent example of a stateless application or process is a web search with a request, an execution, and a response. Since stateless applications and processes are independent of a preservation of session context, they are relatively more fault tolerant, and lightweight in terms of complexity, easier to scale, simpler to migrate across different platforms, and have a much lower relative start and stop response latencies.

While stateless applications and processes are relatively simpler, and suitable for a variety of emerging services that do not require a persistence of context, considerations to support stateful applications and processes is required. The use of containers, which are inherently ephemeral, is a well-suited vehicle for stateless applications and processes. On the other hand, evolving stateful applications and processes to utilize containers and adapt to a stateless environment incur some of the following challenges:

- Ensuring data persistence beyond the lifecycle of the container
- Guaranteeing that the data follows containers whenever they are altered by the orchestrator
- Appropriate data protection
- Controlling the access to data for multi-tenancy and data protection

The stateless VNF in a service-based framework is an easy to manage and deploy function, which utilizes a shared data layer for structured data stored in the Unified Data Repository (UDR) [4] and with the unstructured data storage handled by the Unstructured Data Storage Function (UDSF) [4]. These NDL considerations, in a service-based framework are relevant for applicability in MEC environment data, or for data associated with third-party applications, especially for unstructured data.

The NDL layer is a functional building block, which can be distributed to support a variety of different deployment scenarios, while also satisfying low-latency service requirements. Ultra-low-latency service demands at the network edge, typically require data persistence in close proximity to a run-time container. In such cases a UDR or a UDSF instance can be deployed in close proximity to the run-time workload or even colocated on the same physical server, based on placement and affinity rules.

5.4 Data Placement and Affinity Rules

Data storage policies are used for grouping similar types of data, where the grouping of data is based on the nature of data in terms of high availability, latency bounds, consistency in terms of a single point of

provisioning, and proper level of security. These data storage policies are referred to as data affinity rules. Configurable data storage policies are utilized for rendering appropriate data affinity rules, which in turn promote both a flexible as well as an efficient scheme for diverse data storage, facilitated by the NDL layer [16].

The configurable data storage policies may also dictate the nature of data storage, such as for example, local versus geo-distributed, redundant, asynchronous versus synchronous etc. Additional configurable aspects may include criteria, such as the type of data storage (e.g. embedded memory, flash, or hard disk drive).

The NDL serves as a single point of provisioning for data consolidation and consistency. The data can be shared among different applications, where the associated VNFs share structured data (e.g. data structures that are well-defined in [4]), and where VNFs are associated with unstructured data (e.g. undefined data structures that have no recognizable pattern, such as mobile device context etc.). The data can be exposed to the different clients or applications through different arrangements of a service-based framework.

5.5 Usage Scenarios

Since the NDL provides a unified fabric of provisioning for the entire network control plane, the relevant information is logically stored, provisioned and automatically updated. Since the network related data is consistently collected, consolidated, and stored in one place, data storage requirements are optimized, resulting in a reduced data storage capacity cost.

Furthermore, this consolidation of network related data enables rich analytics to expand and improve the quality of service and the related experience for customers. The vast amounts of diverse data generated from a highly distributed and heterogenous end-to-end system is expected to be an assortment of unstructured and structured data. Autonomic networking requires a leveraging of the various types of data for effectively and efficiently orchestrating, assuring, and fulfilling the service quality and experience demands over corresponding instantiations of end-to-end network slices. The NDL supports autonomic behaviours and related intent-oriented [17] optimization, with the diverse requirements of a multi-tenant, virtualized end-to-end system. Examples, of a few usage scenarios include:

- Analytics
- Orchestration
- Assurance
- Fulfillment
- Autonomic networking

6 ARTIFICIAL INTELLIGENCE (AI) AND MACHINE LEARNING (ML)

6.1 General

Flexibility, adaptability, and the support for a variety of deployment and business models, underscore the emerging service paradigm over next generation heterogeneous and diverse radio access technologies that leverage licensed and unlicensed spectrum. In this emerging and complex context ML provides an indispensable enabling capability under the broad umbrella of AI to support the required attributes for a forward-looking and innovative service paradigm, from an end-to-end perspective. This end-to-end perspective spans the core network, radio network, edge network, user equipment, machine-type equipment, service verticals, and a personalization of experience. An intelligent harnessing of spectrum, network slices, virtual functions, computing, connectivity, and storage resources, require autonomous capabilities augmented by ML to effectively manage complexity resulting from the need to optimize flexibility, adaptability, resource utilization, and performance.

This section delves into the various aspects and directions of ML relevant for a next generation mobile and fixed network that are pivotal for enhancing the service paradigm. The supportive and enabling

advancements in the end-to-end framework, in terms of cognitive and learning capabilities are considered through an application of suitable ML models. Broad and prominent categories of ML models include supervised learning, unsupervised learning, and reinforcement learning. Federated learning [25] allows for cooperative and collaborative learning across different domains using different ML models.

Use cases are delineated to exemplify the relevance and applicability of ML. Some of the prominent categories of examples, where an application of appropriate ML model is necessary for automatic optimization through cognition and learning, include spectral efficiency, inference of an adaptive or predictive action derived from radio interference conditions, capacity and coverage management, energy management, fault mitigation, interaction with Verticals, dynamic network slice orchestration, assurance, fulfillment etc.

6.2 Broad categories of AI and ML

Along the emerging directions of autonomic networking, embedded cognitive capabilities are realized in terms of the various modalities of AI and ML. An autonomic subsystem enables an automated optimization of core and radio network resource utilization, while simultaneously offering an optimized service experience for end-users across human-machine interfaces, and an optimized performance for machine-machine interfaces. These types of optimizations are realizable in a flexible, adaptive, and dynamic manner, to suit the diverse KPIs that characterize the 5G service categories and use cases. The cognitive capabilities of AI and ML offer a variety of use case aligned, and extensible methods to meet the connectivity, coverage, and service demands of a virtualized, decentralized, distributed architectural context [20].

AI and ML are broadly viewed as a class of computer and algorithm assisted intelligence modalities that mimic human intelligence at a task level, with varying degrees of problem complexity characterized by analyzing a given set of data or observations, while determining an optimized solution to meet a desired objective. DL (Deep Learning) is an augmentation of ML rendered through the use of multilayer neural network algorithms to flexibly handle a diverse array of complex use cases, associated with structured or unstructured data Fig. 10. below represents a generalized classification of AI, and ML.

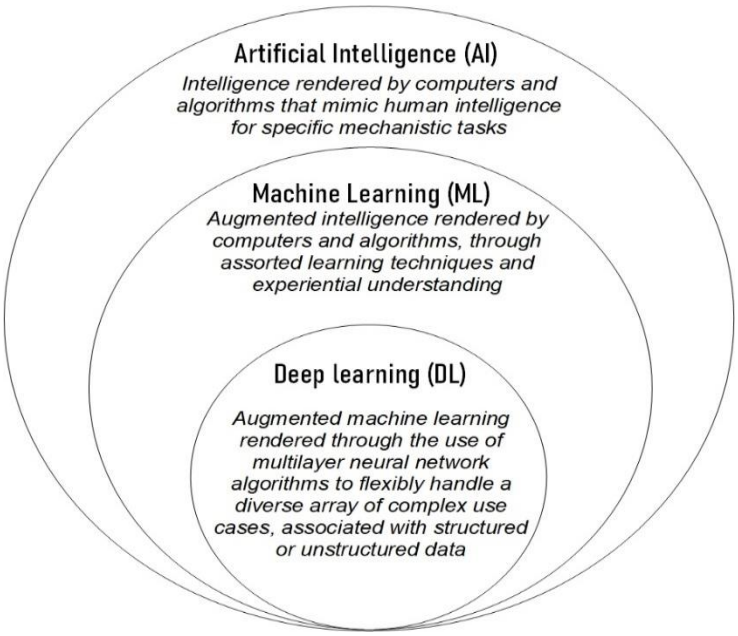


Fig. 10 : Ontology of AI and ML

The functions across any layer of the core network, edge network, transport network, radio access network, or the user equipment are potential reference points to serve as a source of data or as a target for control,

behaving as an input or an output respectively for a given ML function that optimizes the operation of an associated entity in the 5G system. The application of a specific learning model, hinges on the nature of a given entity or feature being optimized in the 5G system. A common framework of architectural building-blocks that are technology neutral is beneficial for harnessing a given ML function and its related interfaces, for technology-specific realizations. Examples of technology-specific realizations include edge computing, fixed-mobile convergence, beam correspondence, spectrum and network resource utilization, service experience and personalization, autonomic management, and control, among several other forward-looking and emerging aspects, across eMBB, mIoT and URLLC categories of service offerings.

The potential of ML capabilities realizable through a variety of learning techniques is significant in terms of managing complexities and harnessing opportunities in the 5G service paradigm. The ML learning techniques are broadly classified as shown in Fig. 11.

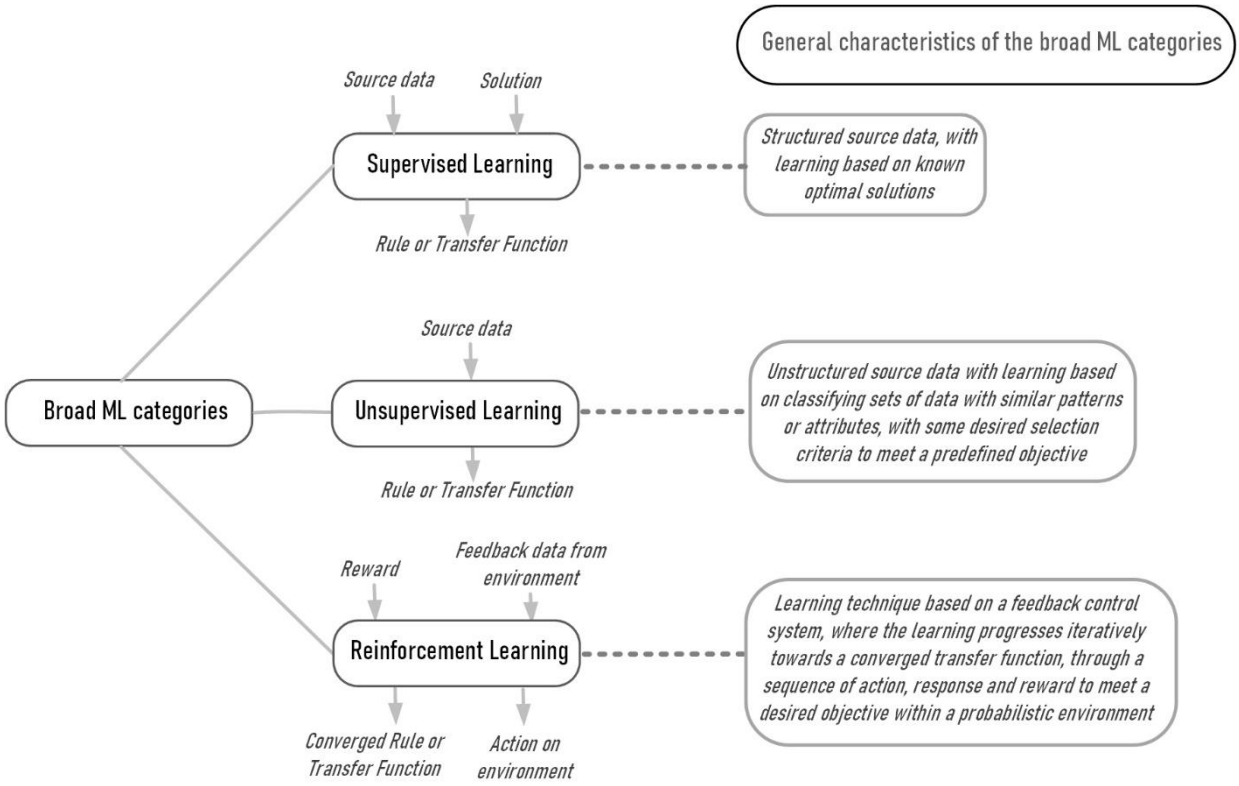


Fig. 11 : Broad classification of ML and general characteristics

The three broad categories of ML as depicted in Fig. 11, are amenable for optimizing a variety of use cases in terms of service KPIs, system performance, resource utilization, energy management, the quality of service experience, and for a variety of innovative and forward-looking scenarios. As referenced in Fig. 11, structured and unstructured data are generalized as having the following attributes:

- **Structured data:** Consists typically of formatted and easily readable quantitative information, such as names, geolocation, address etc.
- **Unstructured data:** Consists typically of information with no categorizable pattern that can be easily read, analyzed, or processed, such as mobility context, audio, video, satellite images etc.

A few examples of use cases suitable for each of the three broad categories of ML include the following:

6.2.1 Supervised Learning

This ML category is useful for utilizing structured data and a known pattern to learn a suitable rule or a transfer function.

6.2.1.1 Examples of applicability

Examples of an application of this ML category, include optimization for a variety of radio layer aspects, such as channel estimation, handover, signal processing, spectrum utilization, among others, where a known model is available. This type of learning is also relevant for upper layer aspects, such as quality of service, end-user behaviors, mobility enhancements, among others, using established models of discovery and learning.

6.2.2 Unsupervised Learning

This ML category is useful for utilizing unstructured data, where an appropriate model is discovered and inferred heuristically from the data to learn a suitable rule or a transfer function.

6.2.2.1 Examples of applicability

Examples of an application of this ML category, include optimization for a variety of heterogeneous network deployments, coexistence of different radio access technologies, fault detection, anomaly detection, end-user behaviors and preferences, intrusion detection, among others.

6.2.3 Reinforcement Learning

This ML category is useful for learning about an environment, using a feedback control system methodology, where the learning is dynamic and iterative, based on an action, response, and reward sequence to converge towards a suitable rule or transfer function for a given environment.

6.2.3.1 Examples of applicability

Examples of an application of this ML category, include optimization for a variety of probabilistic scenarios, in a wireless and mobile network environment, where conditions are not known a priori. Such examples include dynamic resource allocation for network slicing, channel access conditions, decentralized and distributed resource allocation for edge computing, autonomic networking, dynamic spectrum sharing, beam correspondence, radio-frequency coexistence, radio-frequency interference management, dual connectivity and carrier aggregation, network and user equipment or device resource cooperation and coordination, among others.

6.2.4 Federated Learning

Federated Learning (FL) allows for a cooperative and distributed arrangement of learning modalities, including reinforcement learning, among others. This type of cooperative and distributed learning modality enables flexible levels of customization, as well as personalization at the resolution of every end-user for a given service, while scaling effectively, through automation that facilitates operational, resource, and performance optimization. Cooperative and distributed ML techniques work in a complementary manner with MEC that inherently consists of flexible arrangements of access network resources to optimize the end-user service experience. Furthermore, these attributes enable the enhancement of system availability, reliability, and fault isolation, together with improvements in information integrity, security, and privacy for the end-user.

FL enables each local node (e.g. device) to learn a shared or aggregated model in a collaborative manner, while maintaining the training data on each local node, leveraging the increasing availability of computing resources in each local node. The synchronization of computation among multiple local nodes and a centralized datacentre, is accomplished by FL, via a minimized exchange of the necessary computed results. A secure aggregation method [26] uses encryption for protecting the updates associated with individual devices belonging to different parties or domains

The benefits of FL that leverage a distributed arrangement of the radio-access network subsystem, include the following:

- Diverse deployment choices through the flexibility of distributed arrangements
- Avoidance of congestion and higher signalling overhead inherent in a centralized arrangement
- Leveraging of a shared model for automatically maintaining and updating the training data, while also utilizing the power of localized and distributed computing, storage, and network resources at the edges of an end-to-end architectural framework.
- Synchronization of localized and distributed computing, storage, and network resources, with a centralized data centre that has a broad scope, with a minimized signalling overhead in non-real-time, or near real-time.

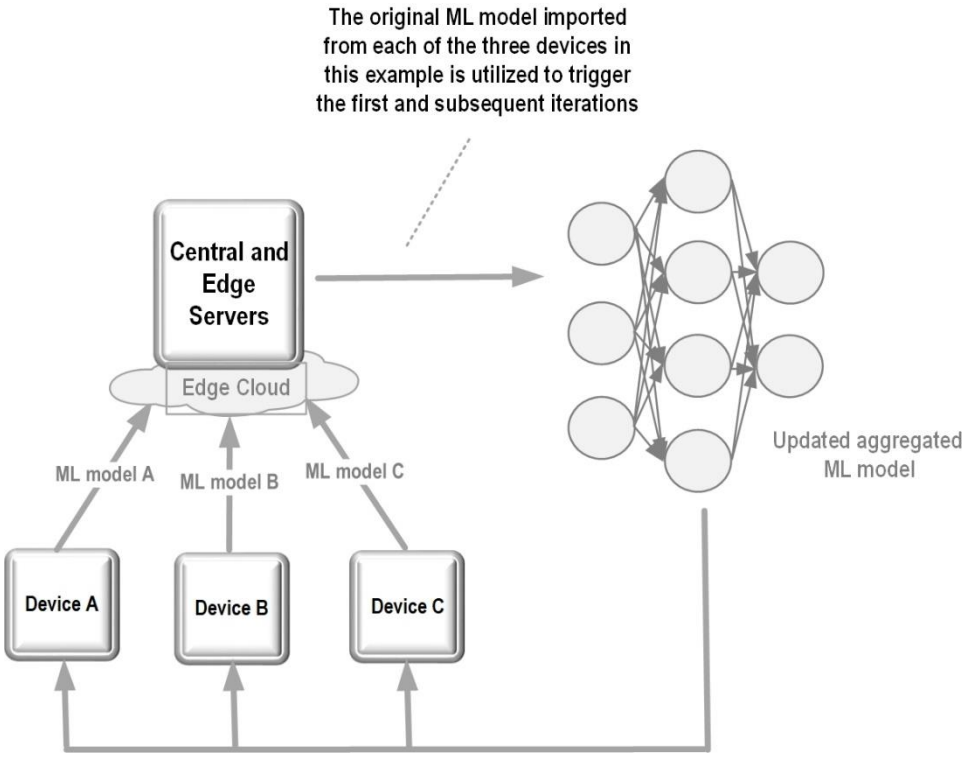


Fig. 12 : Federated Learning (FL) model

A few Illustrative examples that underscore the need for an FL model, as depicted in Fig. 12, include the following:

- Centralized repositories and uploading of user data are avoided, and thereby minimizing a broader exposure of user data, which in turn mitigates privacy concerns, and supports a reduction in security vulnerabilities
- Avoidance of congestion associated with centralization, such as where large volumes of data from a vast number of devices (e.g. connected cars, drones etc.), which in turn leads to a high overhead and higher congestion probabilities over radio access links.
- The challenges associated with data synchronization and integration, as a result of a competitive, security, privacy, or divergent business reasons, in a multi-vendor or multi-supplier, or multi-manufacturer ecosystem of connected cars, where they may not be open to sharing certain data collected from autonomous cars.

Each local or end node manages its own events, resource usage, and mobility behaviour. This local information is utilized to establish a local ML model. Each such local device trains its own ML model using the local information, and then uploads the model updated weight, which is an abstraction of the local raw information, to the centralized or edge server. The centralized or edge server then aggregates all the local model updates, from each device and sends the aggregated model updates to each of the local devices. The process is iterated until a target goal or objective is established. Privacy and security of local device information is retained since the raw information does not leave the local device. This architectural principle has a two-fold benefit, in terms of an improvement in data security and privacy, together with a minimization of communication resource overhead since the weight update data volumes are a fraction of the raw local information volumes. Along with this local information, a system wide vision and behaviour requires to be derived across a multiplicity of devices or user equipment [24].

6.2.4.1 Examples of applicability

The FL model utilizes one or more categories of AI and ML models, for the realization of a variety of forward-looking usage scenarios. Some of the related usage scenarios are identified in this section.

6.2.4.1.1 Network optimization

Among a variety of use cases enabled by FL, a significant use case for an NSP or SP is an optimization of mobile network performance, based on large volumes of data collected from local devices or user equipment.

With the use of FL, the collected raw data from local devices or user equipment is not required to be uploaded to a centralized server over the air, which avoids a wastage of radio-frequency resources, while also enhancing the privacy of user data, by retaining the raw data in the local device or user equipment. FL utilizes the weights of local ML models, which are trained locally, and have a much lower data volume, relative to the corresponding raw information, for delivery to a centralized or edge server.

The centralized or edge server then aggregates the weights from each of the local devices or user equipment and sends back the weights determined by a federated global model, which also gets updated incrementally as the weights of local ML models are processed and aggregated. The federated global model reflects iterative enhancements to each of the local ML models.

6.2.4.1.2 Data privacy across different and cooperating/interacting entities

The FL model facilitates the retention of data privacy, across disparate, cooperating, and interacting entities, while collectively optimizing and automating the behaviour and performance of each of the local devices associated with a federation of entities.

6.2.4.1.3 Application within the UN Sustainable Development Solutions Network (UN SDSN)

The ITU, as a significant organization under the auspices of the United Nations (UN), in the planning, coordination, cooperation, collaboration, and research on a global scale, in terms of telecommunication services, and has been instrumental in providing measures for the UN Sustainable Development Goals (SDGs) [27], with respect to Information and Communication Technology (ICT). The five fundamental aspects of these SDGs include:

- Communications
- Computing speed
- Networking
- Storage capacity
- Computing capabilities

The UN Sustainable Development Solutions Network (UN SDSN) was set up in 2012 to “mobilize global scientific and technological expertise to promote practical solutions for sustainable development, including the implementation of the Sustainable Development Goals (SDGs) and the Paris Climate Agreement. SDSN

works closely with United Nations agencies, multilateral financing institutions, the private sector, and civil society"[28]. The intention is to foster practical solutions for sustainable development, guided by regional SDSNs, through a mobilization of knowledge institutions around the various globally distributed SDGs. These objectives underscore an evolution of ICT over the next decade and beyond. Practical usage scenarios could be envisioned for an application of the FL model, such as for example, the building of an ecosystem of distributed smartphones and IoT devices

6.2.4.1.4 Multi-access Edge cloud performance enhancement

The elements of the FL model enable a collaborative utilization of multiple sources of information, associated with different domains and access technologies (e.g. terrestrial, non-terrestrial, and fixed-wireless), with diverse heterogeneous deployment arrangements, at the network edge. MEC capabilities harness the cloud-native attributes of microservices for flexible multi-access edge network arrangements for the realization of a multi-access edge cloud. The characteristics of multi-access edge cloud inherently provides a variety of choices for the NSP to leverage the requisite levels of distribution for a customization of service experience for human and machine type devices/equipment.

The distributed and decentralized nature of a multi-access edge cloud, is especially valuable for latency-sensitive IoT services, where the edge serving nodes, such as base stations with the requisite computing, storage and transport network resources, in a dis-aggregated and virtualized radio access network, are sufficiently localized, in terms of a service offering and its consumption for an augmented service quality and personalized experience. The localization of computing, storage, and transport network resources in the multi-access edge cloud, allows for a simplification of the resource allocation and decision-making process for an optimized, reliable, and highly available service, while also offering the potential for fault tolerance as a result of the proximity of other neighbouring, cooperative multi-access edge cloud serving nodes. FL modalities are useful for neighbouring multi-access edge cloud serving nodes to have an awareness of resource availability across a cooperating set of neighbouring nodes for optimizing the network performance.

6.3 Architectural enhancements through cognitive capabilities

Autonomic networking, consisting of autonomic management and control, allows for cognitive capabilities that may be arranged in a centralized or in a distributed in terms of DEs (Decision Elements), which are part of a Network Element (NE), or a Network Function (NF). The CMs (Cognitive Modules) that are based on AI and ML algorithms enable autonomic behaviors by leveraging data analytics and feedback control loops for the DEs. The CM may either be shared by one or more DEs or may be embedded within the DE. The centralized DEs, which constitute the Knowledge Plane (KP) interact and interwork with distributed DEs for the policy control of the distributed DEs, contained in NEs and NFs.

The nature and scope of the KP, requires different levels of inference and insight, which could effectively harness the different modalities of AI and ML for system-wide cognition that in turn facilitates self-organization and optimization. The KPs may be arranged to support different levels of network segmentation and distribution to suit specific deployment objectives. A federation of KPs that harness FL models may be leveraged for a cooperative operation of KPs, across collaborating domains (e.g. mobile, fixed-wireless, non-terrestrial etc. types of connectivity.)

The self-CHOP behaviours in an end-to-end service-based framework facilitated by AI and ML provide pivotal architectural enhancements through automation that include:

- Cognitive, self-descriptive, and self-advertisement of NE and NF functions for auto-discovery by other functional entities
- Service orchestration, assurance, and fulfillment
- Dynamic policy adaptation to suit service and business logic
- Integration of cyber-physical systems in a tactile internet [29]

- ❖ Coordination, security, and stability of AI and ML models associated with CMs for consistent and context-aware decision-making
- ❖ Cognitive decision-making aligned with intent-based [17] policies and deployment arrangements

This foundational context provided by autonomic networking facilitates flexible architectural enhancements that embody the required levels intelligence, in terms of AI and ML models, distributed in a system-wide manner for a realization of predictive and reactive behaviors, as demanded by the requirements of a system feature or a service associated with a human or a machine type interface.

6.4 Context for AI and ML Data Model

The stakeholders that leverage the AI and ML data model consist broadly of consumers, over a fabric of network and spectrum resources, partitioned in terms of assorted access networks, transport networks, edge networks, and core networks, as depicted in Fig. 13.

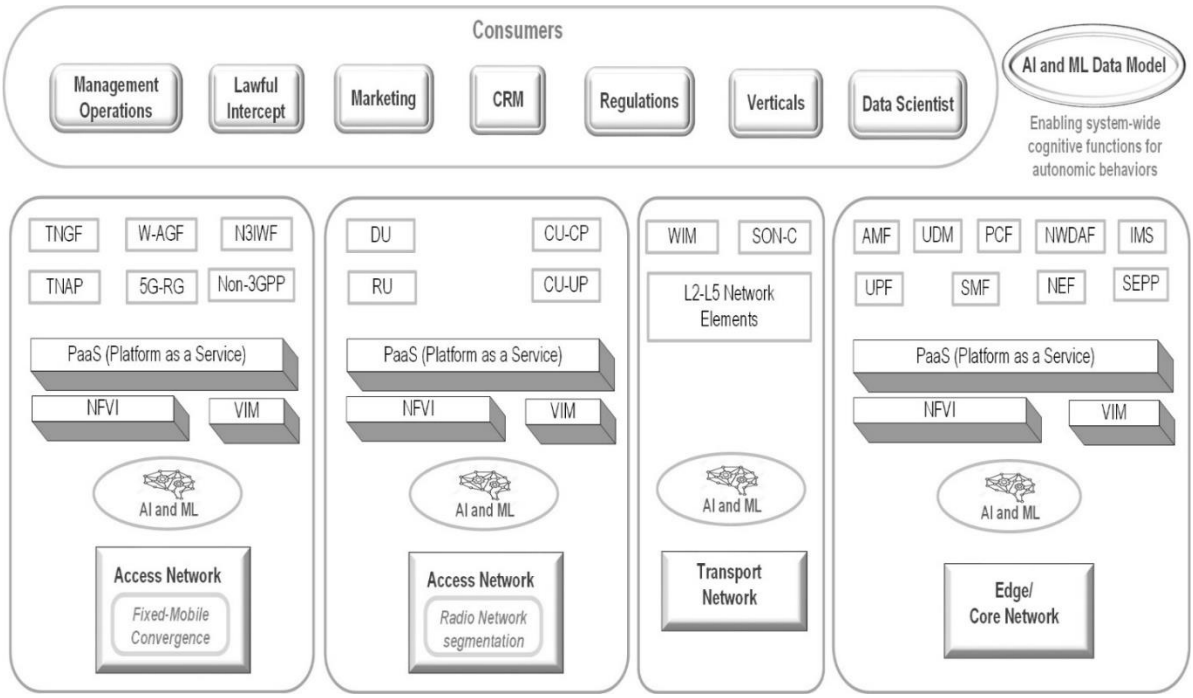


Fig. 13 : Context for AI and ML Data Model

The cognitive capabilities offered by leveraging AI and ML data, underscore flexible levels of information abstraction, through the use of domain-specific or network segment-specific KPs in autonomic networking. This facilitates an effective management of end-to-end system complexity associated with deployment choices and service assurance. Information abstraction enables the articulation intent-oriented self-CHOP behaviours since information abstraction hides the underlying complexity of the supporting infrastructure and resources in a multi-vendor environment.

The AI and ML data model applies to assorted network segments and domains, as depicted in Fig. 13, where the requirements for a given network segment or a network domain are articulated by the corresponding stakeholders. Oversight of these AI and ML models are expected to be provided by data scientists engaged in the optimization and upgrade of these AI and ML models, based on contextual and system environment changes for a maintenance of system-wide self-CHOP behaviours.

6.4.1 Consumers

The consumers of data have consumer-specific requirements that are expected to be a part of the AI and ML model, such as requiring visibility into related data in a given format (e.g. raw, smart data, which is analytics ready meta-data, useful for service assurance). As shown in Fig. 13, consumer types include management and operations, lawful intercept, government and regulators, marketing, data scientist, customer relationship management and Verticals.

6.4.2 Core Network

AI and ML models are leveraged for enabling cognitive capabilities within virtualized functions, associated with the control plane and/or the user plane, for harnessing autonomic behaviours in a service based architectural framework [4]. The enhancement of virtual functions with autonomic behaviours enable self-organizing characteristics, in addition to the Network Data Analytics Function (NWDAF), which collects metrics and performs analysis, such as for managing network slice congestion, load or performance, based on specific algorithms. As the cognitive capabilities evolve in a distributed manner with system-wide scope, as rendered by autonomic networking, these capabilities are anticipated to encompass both physical and virtual functions.

6.4.3 Edge Network

The edge network, based on MEC, shown in Fig. 13 consists of networking, computing, and storage resources that are distributed and closer to the point of interaction with a given service consumer (e.g. human or machine interface equipment). The virtualized resources in the edge network are decentralized and distributed such that the resources required for rendering high-reliability and low-latency services are positioned closer to the network edge, while those services that require intensive networking, computing, and storage resources are handled in the core network. AI and ML models provide the cognitive capabilities for autonomic behaviours in the edge network for optimizing the resource utilization between the core and the edge network [18]. This automates planning, sensing of dynamic service context, adaptation to network conditions, such as capacity and coverage, performance verification, and predictive analytics, in a multi-tenant, and multi-domain environment, for rendering a personalized service experience at the edge of the network.

6.4.4 Access Network

The access network includes terrestrial (e.g. fixed wireless access, mobile access) and non-terrestrial access (e.g. satellite, high-altitude platforms), and represents the boundary of the edge network characterized by radio access. AI and ML models provide the requisite cognitive capabilities, for network slice creation, instantiation, and the life-cycle management of the radio access resources, as part of an end-to-end network slice. The appropriate mapping of radio resources required by a network slice to support a given service KPI, across potentially different radio access technologies and domains is optimized and automated through the use of AI and ML models that facilitate autonomic behaviours.

6.4.5 Transport Network

Similar to the other stakeholders of the AI and ML model shown in Fig. 13, the transport network slice has its own SDN controller like the other components of an end-to-end network slice, between user equipment and applications, which consist of core network, edge network, and radio access network resources. The cognitive capabilities afforded by AI and ML models facilitate the automation, assurance, and optimization of the transport network slice, together with core network, edge network, and radio access network slice components to orchestrate an end-to-end network slice.

The establishment of an adaptive service delivery fabric that responds in an optimized manner, through the cognitive capabilities inherent in a corresponding KP (e.g. associated with the transport network), with closed feedback loops is pivotal for realizing a desired network performance KPI, in the presence of dynamic

conditions, such as faults, errors, failures, threats, service performance degradations, etc., together with changes in workloads, as well as for an autonomic service assurance

6.4.6 Platform as a Service

The cloud-based nature of Platform as a Service (PaaS) enables a virtualized system of resources, where the underlying infrastructure may be distributed geographically across core, edge, transport, and radio networks. The adaptation of broad ecosystem of services over diverse infrastructures is required, in terms of networking, computing and storage resources, hardware configurations, operating systems etc. Furthermore, the consumers of these services in terms of endpoints are diverse and expansive as well, such as mobile devices, sensors, actuators, cyber-physical systems, self-driving vehicles etc. From a business perspective this diversity demands collaboration and cooperation among different actors, such as vendors, developers, public and private infrastructure providers, third-party platform providers, license providers, security, privacy etc.

AI and ML models provide the cognitive capabilities for PaaS to leverage the services across different silos of multiple actors and resources, for a realization of optimized capital and operational expenditure, predictive analytics, network performance, service quality (e.g. machine interfaces), and quality of experience (e.g. human interfaces) [19]. Self-CHOP behaviours afforded by AI and ML models within PaaS, are complemented by CI/CD processes to continuously improve both service quality and service experience. Agile service deployment is among the objectives of a cognitive PaaS to leverage network programmability and virtualization. This demands the use of a variety of virtualization and orchestration schemes for network slicing, service creation, and delivery, over diverse and distributed infrastructure environments. An optimized balance between the granularity of microservices for flexibility, and modularity for ease of configuration, is facilitated through cognitive AI and ML models that automate the optimization of trade-offs between the complexity associated with flexibility, and the rigidity of low granularity.

6.5 Guidance on Test and Certification of autonomics

6.5.1 Generic framework for testing and certifying autonomic functions

In section 4, the requirements for autonomic capabilities are described in terms of a multi-domain, multi-layer autonomic network. To realize the autonomic capabilities of self-configuration and self-adaptation, embodied within the self-CHOP characteristics in a 5G and beyond system, the standardization and testing methods for AI and ML models are required for consistent cognitive capabilities in an autonomic network.

To ease the adoption and deployment of 5G and beyond systems, it is necessary to leverage a test framework for testing multi-layer autonomics, and associated AI and ML algorithms for closed-loop network automation. Fig. 14 exemplifies the context of stakeholders and processes in a test and certification framework for a cognitive autonomic network.

The objective of testing and certification of autonomic functions in autonomic networking is to ensure that the decision-making functions are compliant with the satisfying the KPI targets for cross-domain self-optimization behaviours for resource utilization in the relevant network segments of an end-to-end system. The foundational concept of autonomic networking [9] is described in terms of AI and ML based decision making elements (DEs), LCM (Life Cycle Management) process (e.g. development, training, testing, certification and deployment), with the associated three main stakeholders, namely, AI model based regulator/auditor, third-party AI model based tester, and AI-model based certifier. Each of these main stakeholders provides the related AI support in terms of methodology, function, process, assessment metrics etc., The supporting stakeholders are shown in Fig. 14 [30].

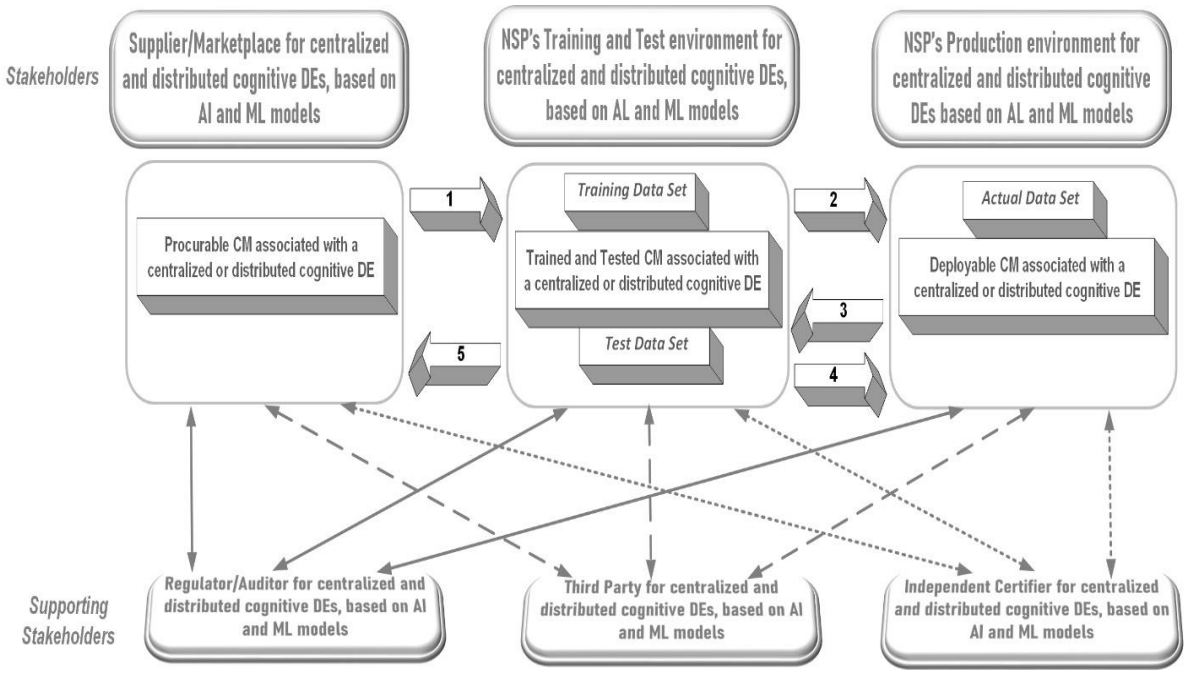


Fig. 14 : Stakeholders in a test and certification framework for a cognitive AMC subsystem

The services offered by these stakeholders in the AI-SS (AI-Support System) may be classified as:

- Regulation as a Service (RegaaS)
- Testing as a Service (TesaaS)
- Certification as a Service (CertaaS)

In a federated deployment arrangement, the testing of AI and ML enabled KPs in an autonomic management and control subsystem [9] is required for the verification and validation of cooperative, coordinated, and consistent autonomic management and control behaviours across the different domains associated with the

Testing Phase / Type of Testing	Validation Phase	Trustworthiness Building Phase	Certification Phase	Testing of Network Deployment Phase	AI model Deployment and Activation Phase	Testing of Network Operation Phase	Testing of Network Optimization Phase
Testing and Validation	x	x	x				
Conformance Testing			x	x	x		
Interoperability			x	x	x		
Integration and User Acceptance Testing				x	x	x	x
Self-Testing						x	x

Table 1: Generic lifecycle testing framework example for AI model based autonomic functions

federated deployment arrangement. A generic lifecycle testing framework example [30] for AI model based autonomic functions is shown in Table 1.

6.6 Usage Scenarios

The AI and ML capabilities are foundational for a realization of a variety of cognitive, closed-loop requirements within an autonomic network for the realization of a dynamic and analytics-driven service assurance. This in turn allows the flexibility and adaptability to establish a predictive, proactive, and augmented customer experience.

The cognitive capabilities offered by different categories of learning models embodied in AI and ML algorithms, form the required ingredients of autonomic functions of the Decision Elements (DEs) in an autonomic network. The DEs facilitate a closed-loop environment for cognitive, self-CHOP behaviours within a given system of network nodes and external interactions, where the self-CHOP behaviours are partitioned into relatively slower and faster control-loops [46].

6.6.1 Autonomic Networking

The cognitive capabilities from an end-to-end perspective, is realized through autonomic networking containing AI and ML enabled functions that leverage appropriate feedback control loops iteratively for establishing an optimized network performance and service quality. The cognitive capabilities among a variety of system-wide benefits of self-CHOP behaviours, also automates the network slice lifecycle management of converged and heterogeneous radio access environments.

Some of the prominent usage scenarios of autonomic networking include programmable traffic monitoring for network slice service assurance [21], and generic framework for multi-domain federated KPs for end-to-end autonomic security management and control [22] for network slicing. A few other examples of AI and ML usage scenarios are also described in this section.

6.6.1.1 Fault detection in the FTTH network

There several usage scenarios for the application of AI and ML models for a realization of optimization and self-CHOP behaviours on a system-wide basis, based on business and deployment objectives. An exemplification of an AI and ML model usage scenario for a cognitive fault detection and optimization, in the FFTH network (i.e. gigabit capable Passive Optical Network (PON)), which may also be applied on a system-wide basis, is shown in Table 2, based on TM Forum usage scenarios.

For this usage scenario example, there are three main considerations:

- Resource management (e.g. network resource, service, or customer management)
- Self-diagnosis
- AI and ML model roles:
- System knowledge presentation for supporting a data scientist
- Decision making support for human in the loop.
- Fully automated decision making, through a closed control loop, as well as for knowledge creation

Use Case Categories	Resource management	Self-diagnosis	Decision support
Use Case scope	This Use Case deals with fault management of fibre access network. It is of potential interest for cities deploying FTTH gigabit capable PON systems or planning to do it.		
Problem statement and Use Case objectives	The goal is to provide an accurate diagnosis capability for FTTH PON networks and to identify the possible root causes of network faults for assistance on site if human intervention is needed.		

Expected benefits and outcomes	❖ Decrease the rate of non-identified faults ❖ Facilitate maintenance of FTTH diagnosis tool through machine learning ❖ OPEX reduction: with an Accurate diagnosis, technicians will spend less time on site to investigate the root cause
Current operational situation	Current tools, for example, in the case of fixed access (copper + fibre) network diagnosis, are proprietary in nature. There are two major drawbacks with a proprietary FTTH diagnosis tool that requires human intervention ❖ Some complex fault configurations are not covered ❖ Updating the FTTH diagnosis tool requires considerable manual intervention
Technical description	The solution should precisely describe the path that leads the observations to the root cause(s) in order to provide an accurate diagnosis of the fault, and subsequently guide human intervention on site if needed.

Table 2: AI/ML mediated fault detection, prevention, performance enhancement in the FTTH network

6.6.1.2 Service Based Architecture

The virtualized and flexible nature of a service-based architecture accommodates a diverse and dynamic topology, such as multi-access edge cloud, network slicing, PNI-NPN (Private Network Integrated – Non-Public Network). The service-based architecture consists of virtual network functions that may be realized as microservices for enhanced granularity, flexibility, service integration.

The NWDAF [47] serves as a logical function in the service-based architecture framework for data collection and analysis. This virtualized infrastructure, with its inherent flexibility, and the awareness of a wide variety of services associated with human and machine interfaces, an adoption of cognitive capabilities in the NWDAF and in the other virtual network functions is required from a system-level perspective to effectively manage complexity. Different types of cognitive AI and ML learning models, based on the role of a virtual network function, would provide autonomic capabilities, such as an automatic management of traffic congestion, network slice configuration, an optimization of system performance, and a variety of other value-added advancements.

6.6.1.3 Network Slice Service Assurance

Different network slices are required to dynamically support a corresponding variety of emerging services (e.g. Verticals), where a given logical network represented by a network slice may consist of resource (e.g. networking, computing, and storage) allocation across different domains, and network segments, such as the core, edge, transport, and radio network. These different network slices that are orchestrated [23] over a shared physical network, have different requirements, in terms of the service quality (QoS) and the service experience (QoE), security, latency, robustness, and reliability, to meet target KPIs associated with services over a human-interface or a machine-interface.

Service assurance is required to satisfy the service level agreements (SLAs), associated with a variety of different network slices and services (e.g. system resource related services, customer-facing services etc.), from an end-to-end system perspective. In this context, AI and ML models are required to automate service assurance through the requisite monitoring and analysis models, where an automatic adaptation to meet the dynamic and diverse network slicing is required for managing complexity and for optimizing end-to-end system behaviour.

7 VIRTUALIZATION IN THE RADIO ACCESS NETWORK

7.1 General

Virtualization decouples software and hardware, which enables NSPs to dynamically pool and scale the computing, storage, and networking resources. This paradigm facilitates an enhanced feasibility in terms of the development and deployment of innovative services to create and serve new markets, as well as to adapt to changing and evolving market demands.

A virtualized Radio Access Network (RAN) enables the transformation of proprietary hardware-based RAN entities, such as the Base Station (BS) into software-based functions executing on high-availability, high-performance, and high-reliability common hardware platforms. This decoupling facilitates a more flexible, agile, and cost-effective arrangements, consisting of a centralized pool of Base Band Units (BBUs), using Virtual Machines (VMs) executing on common hardware (e.g. standard servers), together with Remote Radio Units (RRUs), and transport networks that connect between RRUs and BBUs. The virtualization of the RAN facilitates a separation of the control plane and the user plane, which underpins flexible arrangements of decentralization and distribution.

The separation of the control plane and the user plane in the RAN facilitates higher levels of granularity in terms of network decomposition, with the opportunity to share the data plane, facilitate stateless models, and allow a disaggregation of the RAN in terms of a Centralized Unit (CU) and a Distributed Unit (DU). These flexible arrangements of the RAN can be combined with Continuous Integration (CI) and Continuous Deployment (CD) strategies, together with the cloud-native constructs of a container-based PaaS. Such a framework is significant for enabling the deployment of Network Functions (NFs), and the automation of processes in the system, such as configuration, scaling, migration, anomaly detection, self-organization, decommissioning, energy saving, etc., through the use of autonomic self-CHOP capabilities.

The principles of network slice creation and instantiation in the RAN follow those in the CN to suit architectural arrangements that are pivotal for rendering services over MEC promoting significant advancements in the quality of service experience for the end-user. The virtualization of the RAN, together with the functional virtualization of the CN, enables an end-to-end network slice creation and instantiation to provide flexibility for supporting services with different requirements of performance, capacity, latency, security, reliability, and coverage. The use of self-CHOP capabilities complements a realization of the benefits of CI and CD and open APIs to suit flexible business and deployment models. The self-CHOP capabilities are embedded with AI and ML capabilities for autonomous decision-making that automates the operation of the RAN, and optimization of system performance.

7.2 Architectural Considerations for Split Centralization and Distribution

Flexible architectural options for the RAN, enabled through virtualization, allow various levels of a split between centralization and distribution. A centralized pool of BBUs consist of virtualized functions, while non-virtualizable functions are resident in RRUs. This flexibility requires a high-bandwidth and low-latency fronthaul between the BBU and the RRU.

The decomposition of the RAN through a partitioning of the protocol layers between the BBU and the RRU in addition to the use of virtual functions promotes augmented flexibility and performance trade-offs. The partitioning options between the BBU and the RRU allow for a variety of considerations, such as, diverse deployment choices, KPI requirements (mainly latency), link bandwidth, complexity, centralized aggregation (higher or lower layer split) and capability integration, distribution flexibility, multi-vendor inter-operable interfaces etc.

Fig. 15 shows two prominent split options (with sub-groups) for a disaggregated RAN.

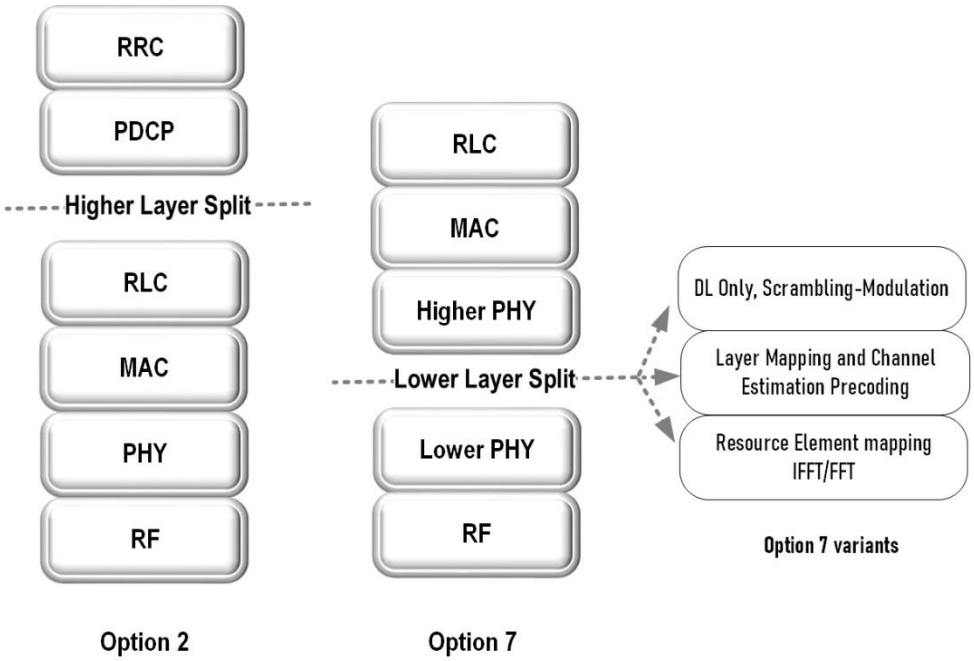


Fig. 15 : Prominent disaggregation options

The trade-off between 'cost and complexity' versus 'latency and transport' requirements pertaining to the different options are generally understood, while requiring specific choices to suit any given set of use cases. While a Higher Layer Split (HLS) offers the benefit of low cost and complexity for implementation, a Lower Layer Split (LLS) offers a higher centralization gain. The LLS option demands a low latency and high bandwidth fronthaul link. For instance, the Option 2 arrangement in Fig. 15 allows a maximum one-way latency up to 10 ms., with a relatively low fronthaul bandwidth requirement (~300 Mbps), whereas an Option 7 arrangement requires a more stringent latency of the order of ~0.1 ms., and a much higher bandwidth (1+ Gbps).

Given that different applications and deployment scenarios demand diverse latency and bandwidth requirements, a disaggregated and virtualized RAN architecture offers the flexibility to support these diverse demands through different RAN functional split options. While there are several RAN functional split options depending on the protocol layer of partitioning between BBU and RRU, Option 2 and Option 7-2x are prominent, in terms of development, interoperability testing and adoption [41].

7.2.1 Trade-offs associated with different split arrangements

A realization of virtualization in the RAN, aligned with virtualization capabilities in the CN, is a significant departure from the existing implementations in a traditional RAN, where the implementations are proprietary in nature with a tight coupling of hardware and software functions resulting in limited flexibility. These flexibility limitations are restrictive in terms of deployment arrangements and multi-vendor choices for NSPs. The decoupling of the software and hardware, provides the flexibility for NSPs to leverage a suitable functional split option, based on software changes, using the same hardware platform. Virtualization in the RAN is pivotal for the realization of network slicing from an end-to-end perspective

A list of requirements for virtualization in the RAN is provided in the following sub-sections.

7.2.1.1 High fronthaul bandwidth

In the case of a centralized RAN arrangement, where the functional split is moved towards the physical layer or between the physical layer and RF layer, a high fronthaul bandwidth capacity is required to establish higher levels of radio network coordination, between the BBUs and the highly distributed RRUs to meet the stringent low-latency KPI.

This requirement reveals that the cost to provide a high fronthaul bandwidth capacity will be proportionately higher relative to the level of distribution. Consequently, the type of functional split requires to be aligned with the latency and RAN distribution demand KPIs to suit a specific NSP service offering or a class of service offerings.

7.2.1.2 Different Functional Split options: Centralization versus Distribution

In the case of latency sensitivity combined with high-reliability use cases, such as the URLLC category of services, which demand a round-trip latency limit of 1 ms. gNB functions are best placed close to devices such as by a colocation of the RRU to minimize the latency. This integration of the gNB functions closer to end users offers a greater extent of distributions, with local breakout choices to application servers for minimized latencies. This latency minimization is bounded by physical distance and the bandwidth of transport between the distributed BBUs and RRUs. Local-breakout of the user plane for device to device communications, while meeting the stringent service KPI demands enhances the service experience through an extension of the edges of the RAN closer to the end-user. Virtualization, from an end-to-end perspective is applicable, where a given network slice would consist of both virtual and physical resources.

For traditional use cases, such as the eMBB category of services, which does not have an as stringent latency requirement relative to the URLLC service category, a more centralized approach that utilizes the HLS option is suitable. The HLS option provides a centralized BBU pooling gain and better interference management and processing through multi-cell coordination. Furthermore, the HLS option is less demanding in terms of backhaul transport requirement to the core network since it avoids the complexity and cost associated with high-bandwidth distributed backhaul.

Virtualization in the RAN should support functional split options to allow various levels of granularity in terms of the degree of centralization or distribution to suit the three main categories of 5G services: eMBB, mIoT, and URLLC. Centralizing larger pools of BBUs, provides mobility over larger coverage areas, while minimizing the mobility signaling overhead. On the other hand, larger pools of BBUs that serve larger coverage areas impair the low-latency demands of services such as video services, interactive gaming etc.

7.2.1.3 Integration and operation complexity

The availability of flexible functional split options promotes deployment choices for NSPs to suit the diverse KPIs associated with 5G services. It is anticipated that the complexity of integrating virtual functions in the RAN in a multi-vendor ecosystem will correspondingly increase, while simultaneously supporting interoperability and flexibility. Hence, the use of an autonomic management and control framework is pivotal for an effective and efficient realization of a virtualized RAN architecture.

7.2.1.4 Open and interoperable interfaces

Disaggregation of RAN through open interoperable interfaces [44] enables the NSP to construct their network in a multi-vendor environment. Together with virtualization, it further provides flexibility, efficiency, agility, and scalability, which are important for the 5G era and beyond when considering further diversifying scenarios and requirements to support services provided by various Verticals.

7.3 Usage Scenarios

Resource sharing in the radio access network requires that the related network slices be able to the support the service associated performance measures, in terms of latency and throughput, over the shared resources of the air interface, the disaggregated fronthaul (e.g. CU/DU split options) and backhaul transport. The sharing of these infrastructure resources, across different domains in a multi-domain environment, facilitates a reduction in the transport network investments. The cloud native nature of RAN virtualization facilitates the sharing of network and spectrum resources, in terms of corresponding network slices.

Examples of network and spectrum resources sharing are described in this section. Virtualization in the RAN leverages SDN controllers in the control layer for a programmable forwarding of information in the data layer. The SDN controllers are logically centralized (e.g. upper control layer towards the core network, with broader scope and lower latency constraints) while they are physically distributed (e.g. lower control layer towards the network edge, with smaller scope and higher latency constraints.). The choice of a level of CU/DU split [42] is dependent on deployment scenarios, business objectives, and the choice of services (e.g. Verticals) that are required to be supported.

7.3.1 Network Resource Sharing among multiple tenants

Network resource abstraction, through network slicing, together with the flexible control layer programmability afforded by SDN, enables multiplexing gain across different network tenants that are supported by a given RAN slice.

The sharing of physical resources, through RAN slicing, allows for a decoupling of the physical resources and their geographic locations. This in turn provides for a logical and flexible partitioning of the physical resources, based on the required networking, computing, and storage resources required by a given RAN slice to support a service KPI. The programmable nature of a virtualized RAN reduces the time-to-market for emerging services, while optimizing performance and resource utilization. From a system-wide, perspective, multiple tenants, such as NSPs, SPs, application/content developers etc., are supported through the capabilities offered by RAN virtualization that leverages resource sharing.

7.3.2 Spectrum Sharing among multiple tenants

Cognitive spectrum resource sharing is pivotal for an efficient utilization of spectrum resources among multiple tenants. These tenants may leverage spectrum databases for a given geographic location, for accessing shared spectrum based on SLAs between the license owner of a spectrum resource and one or more tenants. AI and ML models provide the wherewithal for dynamic spectrum sharing through cognitive capabilities, such as an awareness of geographic location-based spectrum availability, coexistence requirements, and interference management, for a preservation of radio link quality and stability.

For example, spectrum license owner may not utilize the owned spectrum entirely at all times, which would warrant an intelligent spectrum resources sharing between the spectrum owner and a tenant. This benefits the spectrum owner, the tenant, and the service consumer in terms of expanded service availability, and hence enhanced revenue opportunities for both the spectrum owner and the tenant, while at the same time realizing an improvement in the spectrum utilization efficiency.

Spectrum sharing in the low frequency bands (e.g. 600 MHz and 700 MHz) play a pivotal role in coverage enhancements for the NSPs and SPs. The Citizens Broadband Radio Service (CBRS) band (3.55 – 3.7 GHz) [50] is an example of spectrum sharing. Since different NSPs support different service scenarios and are subject to capacity and coverage demands that vary temporally and spatially, such as in the case of carrier aggregation (e.g. frequency band combinations) support, industrial IoT, indoor coverage, integrated access and backhaul, the use of dynamic spectrum sharing is beneficial for improved spectrum resource utilization. Autonomic networking capabilities that are enabled through appropriate AI and ML models, are

invaluable for automating the process of dynamic spectrum sharing across multiple tenants. (e.g. common marketplace opportunities for zero CAPEX service offering enablement).

8 MULTI-ACCESS EDGE COMPUTING

8.1 General

In the emerging and diverse service paradigm an effective and efficient adaptation to customizable service offerings, associated with human and machine type devices/equipment is required. The behavior of an application must be capable of adapting to variations in network resource conditions, such as dynamic changes in capacity and/or coverage, while cognitive capabilities in the network are required to adapt intelligently to preserve a personalized quality of service experience.

Converged radio-access technologies (e.g. terrestrial, non-terrestrial, fixed-wireless) provide the resources for virtualization and distribution of logically distinct sub-networks between the core network and human or machine type devices/equipment. Virtualization of deployment-specific collection of converged radio-access technologies at the Edge Data Network (EDN) [51] is amenable to the use of microservices to realize suitable levels of flexibility and granularity. These attributes enable cloud-native behaviours at the network edge, for realizing a multi-access edge cloud, based on MEC.

The attributes of multi-access edge computing allow for the availability of computing, storage, and networking resources closer to a human or a machine interface to augment the service experience and quality. This strategy alleviates the burden and overhead associated with the utilization of computing, storage, and networking resources that are located in a centralized or a remote cloud. Consequently, the latency associated with resource allocation and the instantiation of a required network slice for supporting a variety of services is minimized through a localized and distributed availability of user plane resources in the multi-access edge cloud.

The cognitive capabilities of the autonomic networking can be utilized for the orchestration, instantiation, and anomaly prevision [53] to maintain and preserve the service quality and experience for services rendered over a multi-access edge cloud. Anomaly prevision allows for the predictability of a malfunction or a fault condition as a result of environmental changes in the system, which in turn would allow for a look-ahead corrective action for preserving the KPIs associated with a given service.

8.2 Converged access and virtualization

At the network edge converged access consisting of terrestrial and non-terrestrial access technologies facilitate the support for services with diverse KPI requirements, as well as ubiquitous service availability closer to human and machine interfaces that promote both service quality and experience. Disaggregation and virtualization at the radio network edge provides flexible deployment arrangements, in terms of the radio baseband and physical layers using interoperable partitioning interfaces.

Virtualization, realized as centralized clouds, is inadequate to meet the stringent and diverse KPIs of emerging services. A variety of emerging services that may be classified as requiring strict latency bounds, highly reliability radio links, low-latency VNFs, and wide bandwidths, together with the requisite networking, computing, and storage resources need to be positioned and distributed in a network edge cloud [52]. A few examples of such services that leverage the resources in distributed network edge cloud, are described in this section.

8.3 Usage Scenarios

The virtualization of the RAN and edge computing are complementary and cooperate [52] to localize the networking, computing, and storage resources for a variety of emerging services that harness network slicing from an end-to-end perspective.

The proximity to human and machine interfaces at the edge of the network, enables the optimization of transport efficiency, which mitigates unnecessary transport links and hops, and simplifies the compliance to security and regulatory policies to contain user data within a localized geographic region for services, while also satisfying system performance requirements. In concert with the cognitive capabilities of autonomic networking, the following are a few examples of usage scenarios that harness edge computing for meeting stringent and diverse KPIs:

- Gaming XR (eXtended Reality):
  - Enablement of a robust and rich visual experience in real-time including immersive VR/AR platforms, cloud gaming for a multi-player environment, over multiple-access technologies

- Manufacturing:
  - Processing of large volumes of data generated in manufacturing can occur at an edge data center for near real-time response (e.g. predictive maintenance etc.), where the sending of data to central data center is infeasible as a result of the limitations of bandwidth and latency bounds
  - Collection, filtering, and aggregation of data from multiple manufacturing machines, processes, and systems to optimize the manufacturing process in real-time and to determine the best course of action through inference based on AI and ML models

- Security:
  - AI and ML model based real-time video analytics and cognitive capabilities, which enable state-of-the-art security solutions for business and public services

- Retail:
  - Enablement of the delivery of an engaging experience based on insights derived from real-time customer behavior, through smart sensors/cameras and AI and ML based analytics

- Automotive:
  - Navigation assistance for autonomous vehicles using real time communication with the network, other vehicles, and pedestrians
  - Other examples of navigational assistance, including the assistance for movement through intersections, traffic congestion alarm, vehicle software update, vehicle health diagnostics, real-time situational awareness, lane change warning, speed harmonization with other vehicles, detection of vulnerable road user, tele-operated driving, parking support, cooperative maneuvers, green light coordination for optimized traffic flow, vehicle platooning etc.
  - Reduction of the transmission volume of data to a remote central data center

- Automation of network management:
  - AI and ML model based network and traffic flow optimization to reduce network resources (e.g. computing, networking, spectrum resources etc.), human and machine interface service quality and experience

- Smart City and Surveillance:
  - Enablement of high reliability for critical connectivity (e.g. drone control)
  - Enablement of connected devices for efficiently and automatically reporting road issues, traffic pattern updating, weather fluctuations etc.
  - Enablement of a smart stadium, through a provision of temporary access to support the necessary bandwidth, a high density of connectivity etc.
  - Video surveillance for law enforcement, and crowd management
  - Dynamic toll/parking pricing

- Automatic monitoring of environmental (e.g. air quality etc.) conditions for making appropriate recommendations

9 DISTRIBUTED LEDGER TECHNOLOGY

9.1 General

The concept of DLT, introduced in [3] consists of a secure, decentralized, and distributed database for accessing and modifying entries associated with virtual or physical resources associated with a given distributed ledger. This concept underscores the capability of not requiring a centralized or third-party mediation. Disintermediation is facilitated through the use of cryptographic distributed data processing techniques, broadly classified as a blockchain protocol. The use of these techniques avoids the overhead of intermediaries for the oversight of transactions between any two entities or among a defined group of entities that share a distributed ledger.

In a distributed ledger consisting of two or more processing nodes, every node maintains a copy of the entire ledger. With disintermediation there is no central authority that oversees or audits the processing node or its ledger copy. Each transaction between participating processing nodes is recorded in the distributed ledger as a block. Every subsequent block in the distributed ledger, resulting from a new transaction, contains the hash value of the previous block. This promotes the data integrity of the distributed ledger. The distributed ledger grows as a linked-list of data blocks or a chain of data blocks referred to as a blockchain.

Every data processing node that needs to access, and update a given distributed ledger is required to meet the established criteria associated with the distributed ledger. These criteria are classified under a consensus scheme configured for the distributed ledger. Among several consensus schemes, the proof-of-work consensus protocol utilizes a nonce to create a cryptographic hash that meets the criteria such as, one-way bias[1], pseudo-randomness[2], collision-resistant[3], and deterministic[4]. The level of difficulty associated with creating a nonce is based on the criteria required for producing a hash value. Once such as nonce is established by the data processing node, it is then able to generate a data block and broadcast it to all the nodes associated with the distributed ledger. In such a proof-work scheme the peer nodes accept the longest chain to allow the growth of the distributed ledger as new transactions are recorded.

The challenge with such a proof-of-work scheme is that peer data processing nodes with greater computational speed or hash rate calculation potential, has a higher probability of generating a new block in the distributed ledger. Consequently, if such a data processing node has a computation speed that exceeds the total capacity of all the other peer nodes, associated with the distributed ledger, then the integrity of the entire distributed ledger is vulnerable [32] [33]. This is referred to as the fifty-one percent attack scenario and is a known drawback of a permissionless distributed ledger, which allows maximum flexibility for any public data processing entity to perform transactions, without an a priori identification.

With this context the permissioned type of distributed ledger is relevant in an end-to-end next generation framework. This avoids the computational overhead and the security vulnerabilities of a permissionless distributed ledger. With a known set of distributed data processing nodes, the integrity of data in the distributed ledger is verifiable using simpler consensus schemes. Appropriate levels of trade-offs are to be considered, in terms of flexibility, disintermediation, and the automation of transactions to be performed by a verifiable set of distributed processing nodes.

---

[1] Easy to calculate output from input, while not impossible to calculate the input from a given output.
[2] A change in input should produce an unpredictable output. For example, if the hash value of 3 is 6, then a hash value of 4 should not be 8.
[3] Two different inputs to a hash function, should not produce the same output
[4] A given input to a hash function should always produce the same output

9.2 Smart Contract

The permissioned type of distributed ledger holds the potential for minimizing the latencies and the costs associated with data access or exchange. It also incorporates the notion of Smart Contract (SC) [34], which automatically executes configured actions and service level agreement without intermediaries, while providing near real-time evidence of any tampering, which in turn provides a framework for compliance with relevant policies or a regulatory regime.

The self-CHOP capabilities facilitated by autonomic networking promote opportunities for automatic end-to-end system optimization, in a virtualized leveraging of networking, computing, and storage resources, where these resources may be within the same NSP or SP domain or across multiple NSP and SP domains (e.g. RAN sharing, resource sharing for Verticals, roaming, common marketplace etc.). The sharing of resources requires the appropriate compliance with bi-lateral or multi-lateral SLAs, where more than a single NSP or SP is involved. The use of SC promotes the automation of SLAs among NSP and SP domains in a distributed, secure, trusted, privacy-protected, sovereign (e.g. identity ownership), and autonomous manner.

The SC resides within a distributed ledger, and consists of a collection of functions and data, which execute automatically for compliance with a given SLA or a contract associated with NSP and SP domains, participating in the given distributed ledger. The data in the distributed ledger may be modified or probed by calling functions pertaining to the SC, or by initiating an interaction with the SC. The functions in the SC execute automatically in response to an SC function call or an initiating interaction.

Collaboration across multiple NSP and SP domains, using an SC within a given distributed ledger obviates third-party brokerage overhead, while optimizing performance over virtualized, distributed and SLA conformant resource sharing, to satisfy service related KPI targets, spectrum sharing, coverage, latency, capacity, and service experience related to an emerging Vertical ecosystem (e.g. Industry 4.0 [39], tactile internet, e-health, autonomous vehicles etc.). The SC augments the creation and delivery of services through the enablement of transparency and traceability of data shared and exchanged among multiple NSPs and SPs in the distributed ledger.

9.3 DLT and MEC

The capabilities offered by a distributed ledger support the benefits of localization and distribution of network, computing, and storage resources that are necessary to support MEC services. At the network edge, DLT consists of a distributed network of functions that process, validate, and store changes in the system (e.g. service, business, and user transactions etc.). These changes in the system are replicated, updated, and maintained as records in the distributed ledger.

Since the computing nodes participating in the blockchain represent a distributed ledger, there is no centralized trust authority. The shared trust is maintained among all the participating nodes for a given distributed ledger, based on a collective consensus scheme to validate any transaction associated with the distributed ledger. In this manner DLT offers distributed security, and privacy for MEC services, with high availability and reliability, while also complementing the self-CHOP behaviours of autonomic networking at the network edge, in a virtualized service based framework.

With the diversity and expansive nature of MEC services that are likely to have a variety of different access protocols, especially in the IoT space, DLT provides a common and distributed scheme for authentication, which obviates the limitations of scalability and security encountered in the case of centralized architectures for authentication. Among a myriad of diverse usage scenarios, DLT enables the capability for a real-time monetization of MEC services, such as the billing for resources (e.g. different combinations of networking, computing, and storage etc.) [35].

9.4 DLT and Autonomic Networking

DLT and autonomic networking provide complementary and distinct capabilities for service innovation, while leveraging the cognitive capabilities of AI and ML to effectively manage the end-to-end system complexities, while optimizing the end-to-end system performance and the service experience.

These capabilities are foundational for an emerging digital economy. Together they influence the formulation of creative and flexible business models and deployment choices, with an end-to-end system scope. An end-to-end integration of the business-service-network aspects with the operating business model, articulated in the form of various enablers that are standardized, has significant benefits in terms of the following:

- Satisfaction of common objectives of a dynamic and value-generating ecosystem of stakeholders, such as NSPs, SPs, Vendors, Independent Software Vendors (ISVs), Integrators, Data Owners, Data Managers, Customers, and Verticals
- Complementarity of DLT and autonomic networking results from each of the capabilities being mutually supportive, where the key features of DLT (e.g. security, privacy etc.) can be leveraged by autonomic networking, which provides optimization and automation capabilities for DLT use cases.

The emerging diversity and expansion of the evolving service paradigm demand a high level of automation fostered by underlying and forward-looking enablers. DLT promotes the automation of dynamic partnerships and flexible deployment scenarios for the creation and delivery of innovative services, while leveraging the benefits of distributed platforms offered by multiple partners. The latter arrangement allows for a corresponding reduction in the CAPEX investments (e.g. zero-CAPEX in a common marketplace) incurred by individual NSPs or SPs through a multi-lateral partnership. With the demanding and diverse KPIs associated with the main service categories of eMBB, mIoT, and URLLC, the leveraging of capabilities enabled by DLT, in concert with autonomic networking is significant for optimizing service quality, experience and personalization.

9.5 Usage Scenarios

It is anticipated that in the distributed ecosystem of 5G and beyond service paradigm, DLT provides an automated and secure shared database, with a variety of features that can be selectively utilized to suit a plethora of business models, deployment scenarios, and diverse partnerships. In other words, DLT offers an optimization of CAPEX and OPEX, while automating the rendering of forward-looking services in a distributed environment, such as at the network edge. Illustrative use cases include:

- Augmentation or reinforcement of existing infrastructure
- Short-term acquisition of infrastructure to support an event or location (e.g. stadium, mall, city, campus, country, region etc.)

A few examples of a diverse ecosystem of emerging services enabled by DLT in the common marketplace (e.g. blockchain-based data and telecommunications infrastructure [36] are described in the following subsections.

9.5.1 Zero CAPEX model

The reduction of deployment cost, using a zero-capex model, together with a reduction of energy consumption are significant requirements to foster the adoption of an emerging 5G ecosystem. According to estimates [38] the CAPEX is anticipated to increase by around 60% or more in the next three years. The mobile equipment and IoT generated internet traffic volumes are expected to grow rapidly with the increasing levels of service decentralization and distribution through MEC arrangements that promote enhancements towards a personalized service experience. Access network convergence across terrestrial and non-terrestrial systems allow for a variety of deployment arrangements for further augmenting the service

experience closer to human and machine interfaces, across the main service categories of eMBB, mIoT, and URLLC.

Zero CAPEX is achievable through the use of a business model that enables opportunities for new business creation and revenue harvesting, where a typical investment in the deployment of infrastructure is not profitable. Examples of such usage scenarios include the following:

- ❖ Reinforcement of existing infrastructure
- ❖ Temporary acquisition of infrastructure to offer services for an event in a geographical coverage area, or location, such as a country, where the NSP does not own the infrastructure, for expanding their services to new customers, thereby increasing their coverage footprint

The challenges associated with a fulfillment of these usage scenario examples is to craft an efficient and straightforward scheme that does not leverage the traditional and cost inefficient 'Make/Buy/Rent' models of mobile network resources and infrastructure. These legacy models are based on long-term, relatively rigid commitments and shared procurement rules that are well-established, with trust and confidence built over years.

On the other hand, the shift towards new NSP business models that harness an auction-based, on-demand sourcing of infrastructure and resources, supported by the emerging model of a marketplace, with the requisite levels of trust, confidence, transparency, traceability, sustainability, and compliance with regulation and legislation, provides a viable pathway for satisfying the usage scenario examples. Agile collaborative modalities are an integral ingredient of such NSP business models for the exploration of expanding business opportunities and harvesting related revenue streams.

9.5.2 DLT based Common Marketplace Platform

The common market place platform is enabled by DLT and provides a secure distributed environment for engaging multi-domain (e.g. NSPs and SPs) partnerships and collaboration for service innovation (e.g. services over MEC). 0, exemplifies the business impacts associated with the operationalization of DLT based common marketplace platform pertaining to three usage scenario examples with respect to 5G network densification. The table also delineates six common challenges that need to be overcome as motivated by the business impacts, with respect to these three usage scenario examples. The common benefits for the three usage scenario examples are also identified in 0 [36].

<table>
<tr><th>Business Impact</th><th>Description of challenges, usage examples, and benefits</th></tr>
<tr><td rowspan="2">➢ Built-in trust and confidence through consensus, for enabling business model innovation

➢ Enablement of new business opportunities

➢ Improvement of business assurance and real-time transparency

➢ Move towards guaranteed SLA</td><td>Top six common challenges associated with each of the following three usage scenario examples:
• Challenge 1. Continuous Coverage Improvement
• Challenge 2. Spectrum Efficiency
• Challenge 3. Continuous Capacity Enhancement
• Challenge 4. Energy Efficiency
• Challenge 5. Lowering Backhaul cost
• Challenge 6. Deliver Network Slices URLCC, mIOT, eMBB</td></tr>
<tr><td>Three usage scenario examples:</td></tr>
</table>

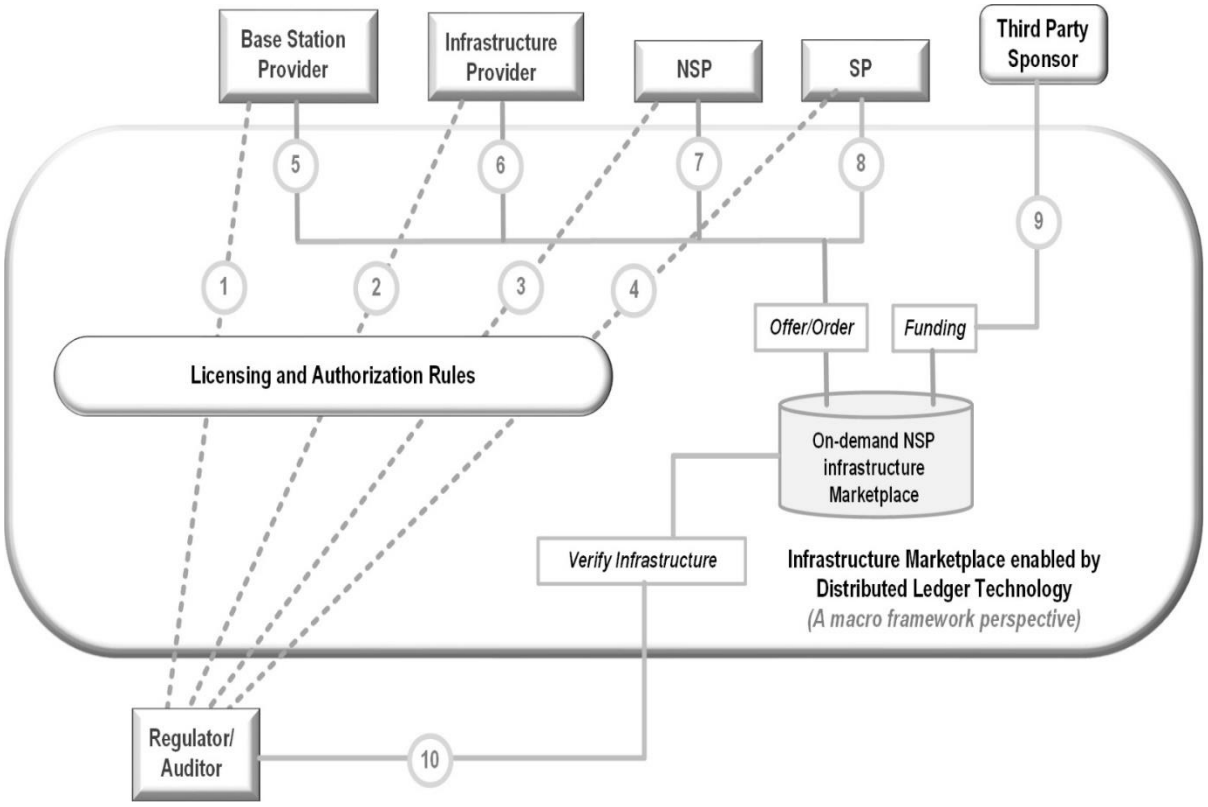
➢ Reduction of disputes, significant reduction of delays, and prevention of revenue leakage as a result of mismatched data	❖ **Usage scenario #1** NSP 5G network densification and/or reinforcement of existing network capacity/coverage ❖ **Instrument** Common marketplace platform for the Telecom Infrastructure ❖ **Usage scenario #2** NSP 5G network densification complemented by on-demand energy as an asset ❖ **Instrument** Common marketplace platform extended to energy as an asset for the Telecom Infrastructure ❖ **Usage scenario #3** NSP 5G network densification complemented by on-demand energy as an asset ❖ **Instrument** Common marketplace platform leveraged for energy collaboration by the Telecom Infrastructure
	Common Benefits • Reduction of reconciliation cost through near real time data availability (e.g. current processes may take about a month for the settlement of reconciliation, and longer for dispute resolution) • Improved customer experience by providing near real-time billing • The common data model of the common marketplace platform, provides generic support for wide variety of usage scenarios • Cost reduction relative to traditional marketplace Real time transparency of data in the network, reduction of the number of disputes between asset providers and NSPs, and reduced settlement times

:

Table 3: Usage scenarios and benefits

The context for a common marketplace platform is depicted in Fig. 16, which is an illustrative example [36] leverages the distributed ledger with the main actors, associated roles, and a demarcation of responsibilities. The main transactions between the actors are classified as follows:

- Non-commercial transactions associated with the auditing and verification of enforcement of regulation/legislation (shown as dotted lines)
- Commercial transactions associated with trading and monetization



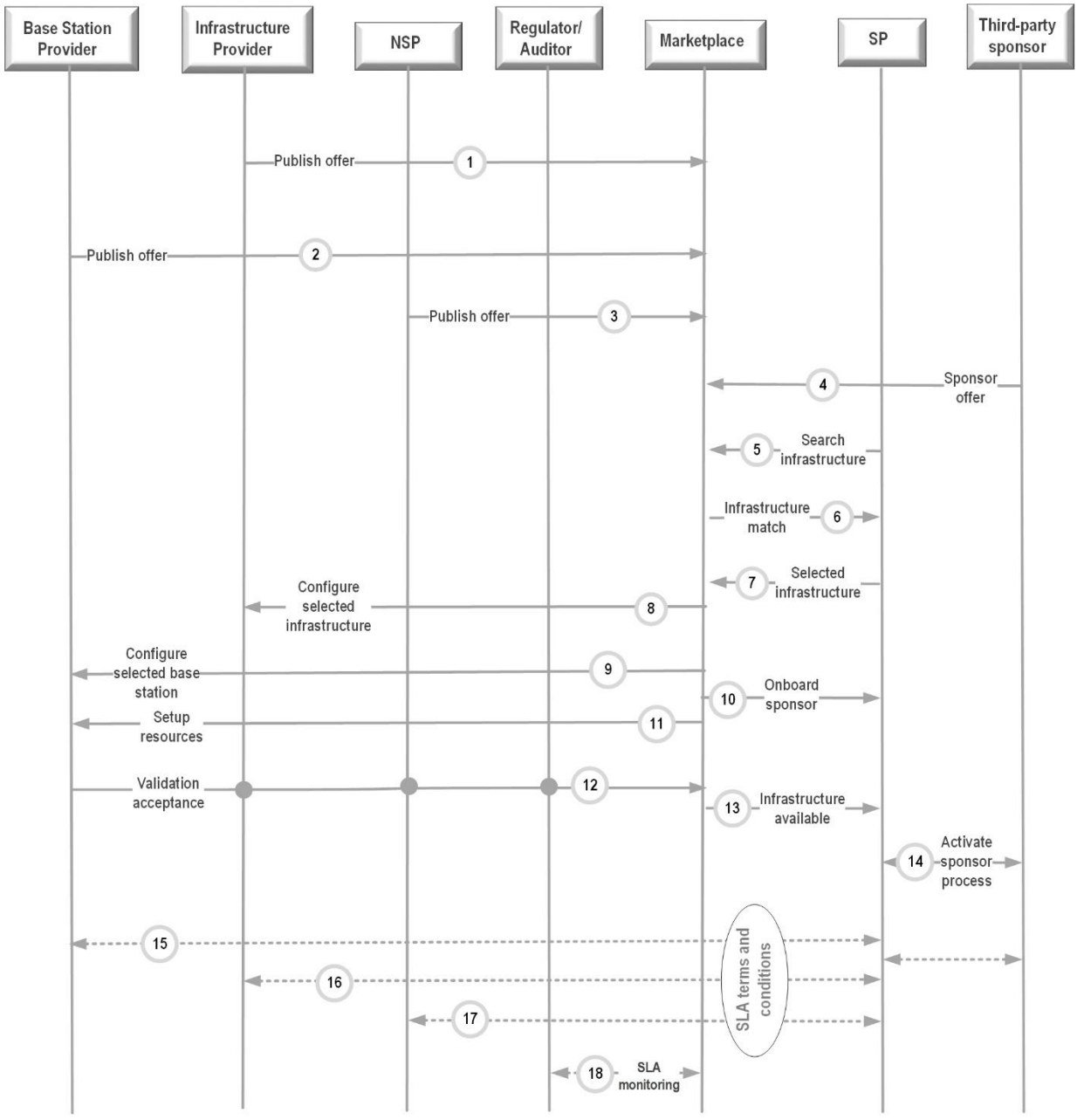
:

Fig. 16 : Common Marketplace context, actors, and interactions

9.5.3 Exemplification of interactions in a Common Marketplace platform

An example of a sequence of interactions among the actors in a common marketplace platform is depicted in 0 [36]. In this example, cooperation among the actors in the common marketplace is harnessed to realize the required networking, computing, and storage resources required to realize a given service offering for a service provider that leverages the common marketplace.

The actors in the illustrative example in Fig. 17 are those depicted in the context of the common marketplace platform shown in Fig. 16.

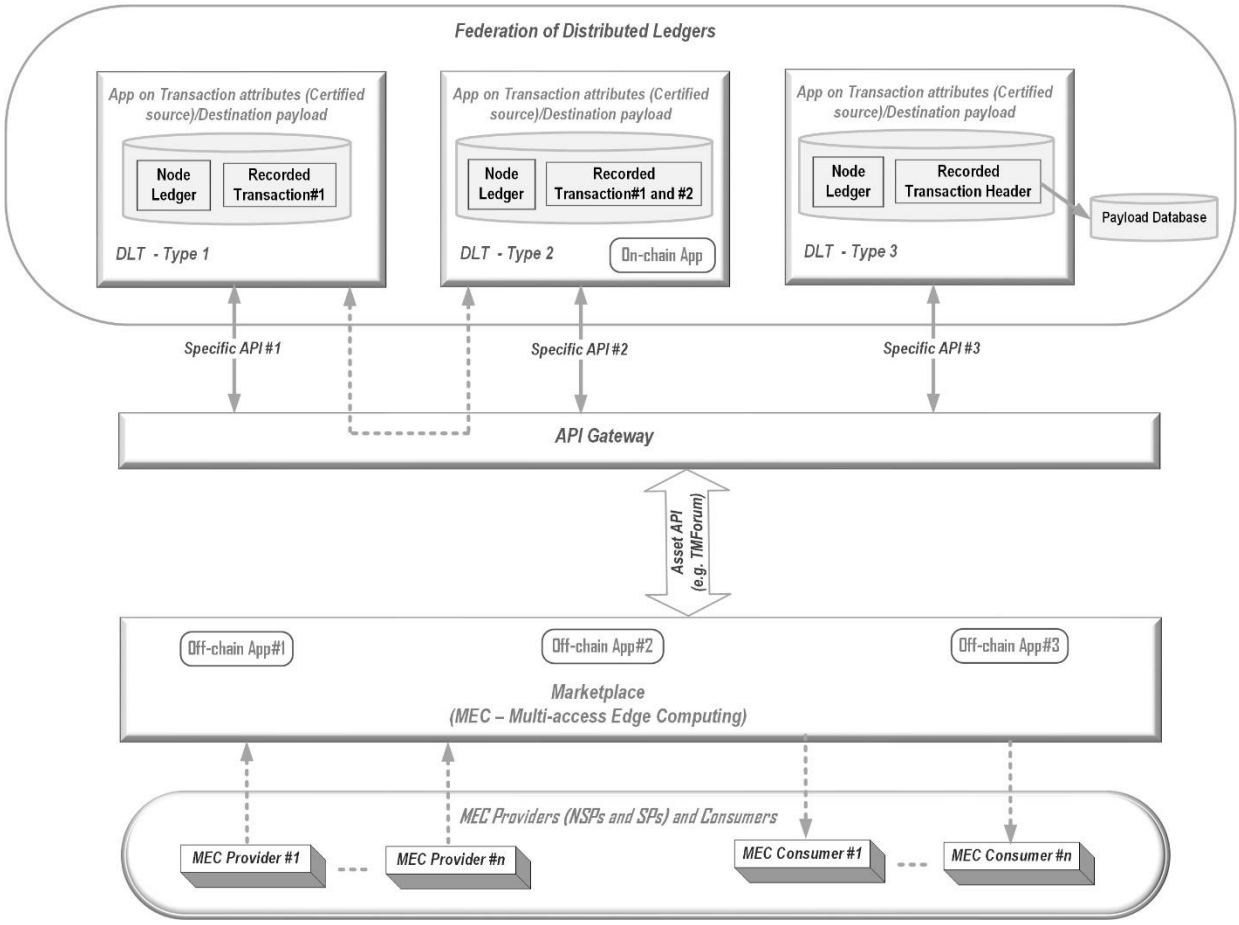


:

Fig. 17 : Example of interactions for service activation in a Common Marketplace

9.5.4 Federated DLT platform in the Marketplace

In the emerging service paradigm, the NSP requires the operationalization a given service offering, aligned with a business and deployment model. The operationalization of a given service offering may require the traversal of multiple DLT platforms, where each provides a necessary ingredient (e.g. spectrum, infrastructure etc.) as part of the composition of the service offering. An illustrative example for satisfying such a business objective and deployment model, which enables interworking across a selected menu of DLT platforms, is to leverage an API gateway that provides an API exposing function, with interoperable interfaces to a federation of distributed ledgers. This is depicted in Fig. 18 [37], for a service offering over a MEC oriented business and deployment model.



:

Fig. 18 : Example of federated DLT platform usage scenario for a service offering over MEC

The main capabilities offered by the use of a federated DLT platform (permissioned or permissionless), in this example depicted in Fig. 18, include the following:

- Traceability of transactions (e.g. timestamped records)
- The trustworthiness and integrity of transactions are validated by consensus, depending on the security level of the considered consensus type or algorithm.
- Authentication of the identity of a “Certification Authority”
- Immutability of a trusted data record

The DLT platform design blueprint characteristics to be exposed and discovered in the marketplace, include the following:

- Payload Type consisting of size, security level, transactions per second (e.g. report/file, token etc.)
- DLT Type consisting of consensus type, with or without payload recording
- Application type for managing transactions:
  - Onchain-based application using data recorded in the DLT platform and validated result, via consensus or via delegation to one or multiple elected node(s)
  - Offchain-based application with or without data validation in the DLT platform

In the illustrative example of Fig. 18, the following describes the main aspects of the federated DLT platform:

- Federated Distributed Ledger Layer:
  - This layer shows three different types of DLT types, differentiated by their respective consensus algorithms (e.g. DLT-Type 1, DLT-Type 2, and DLT-Type 3).
- API Gateway Layer:
  - API gateway ensures interoperability and interworking among the three different DLT types in this illustrative federated of DLT platform model for interaction with a marketplace. In this example the principle embodied in the model is that each DLT platform in the federation interacts with the API gateway, using its own suit of APIs.
- Marketplace layer:
  - The market place layer embeds a variety of off-chain transactions of value, which are outside the blockchain of the distributed ledger, such as onboarding, service orchestration, billing reconciliation and settlement etc.
- Provider/Consumer layer
  - Providers of assets (e.g. MEC networking, computing, and storage resource) expose their assets in the marketplace to be discovered and consumed.
  - Consumers of assets discover the offered assets in the marketplace and request the selection of an available and requested asset to fulfill their requirement. In this example, the assets are associated with the MEC resources to facilitate an NSP to build an edge cloud for service delivery (e.g. URLLC network slices etc.) for NSP customer (e.g. Industry 4.0)

10 VERTICAL MARKET

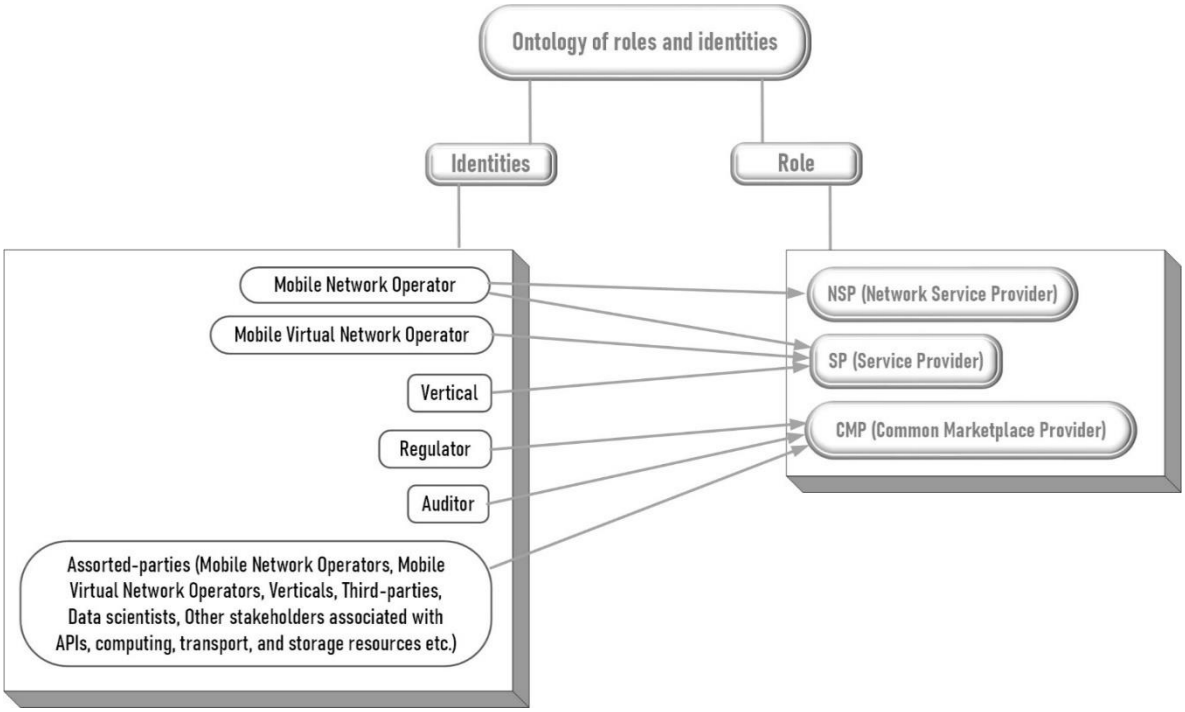
10.1 General

Network slicing is a foundational architectural requirement within the context of an end-to-end framework for a realization of the next-generation service paradigm. This enabling requirement of network slicing requires appropriate resource allocation and utilization within and across different administrative domains. The objective is to facilitate a seamless service experience, whether the service usage utilizes connectivity over a home network domain or over a roaming network domain [45]. A coordination of resources across multiple cooperating domains is required, where network slicing serves as a logical scheme that harnesses a dynamically or statically configured set of distributed resources, to enable service innovation. The enabling capability of network slicing across different domains, while preserving the required service KPIs across the different cooperating domains is essential for the rendering of a consistent service experience for services associated with an expanding Vertical market.

The leveraging of the control plane together with the management and orchestration plane, which enables a flexible and distributed network slicing capability for the data plane, is a critical aspect to meet the quality of experience associated with a high-level of service personalization. This implies that the network is enabled with higher levels of service and user preference discernment, together with an effective management of the associated increases in complexity, which demand cognitive functions to be available in a service based architecture. Autonomic networking serves as an oversight capability for a cognitive system, which enables the requisite levels of automation that match the associated KPIs of emerging services. The service personalization demands are anticipated to be a catalyst for the expansion and diversity of Vertical markets.

10.2 Ontology of identities and roles

An ontology of identities and roles in an emerging and expanding ecosystem of services is depicted in Fig. 19.

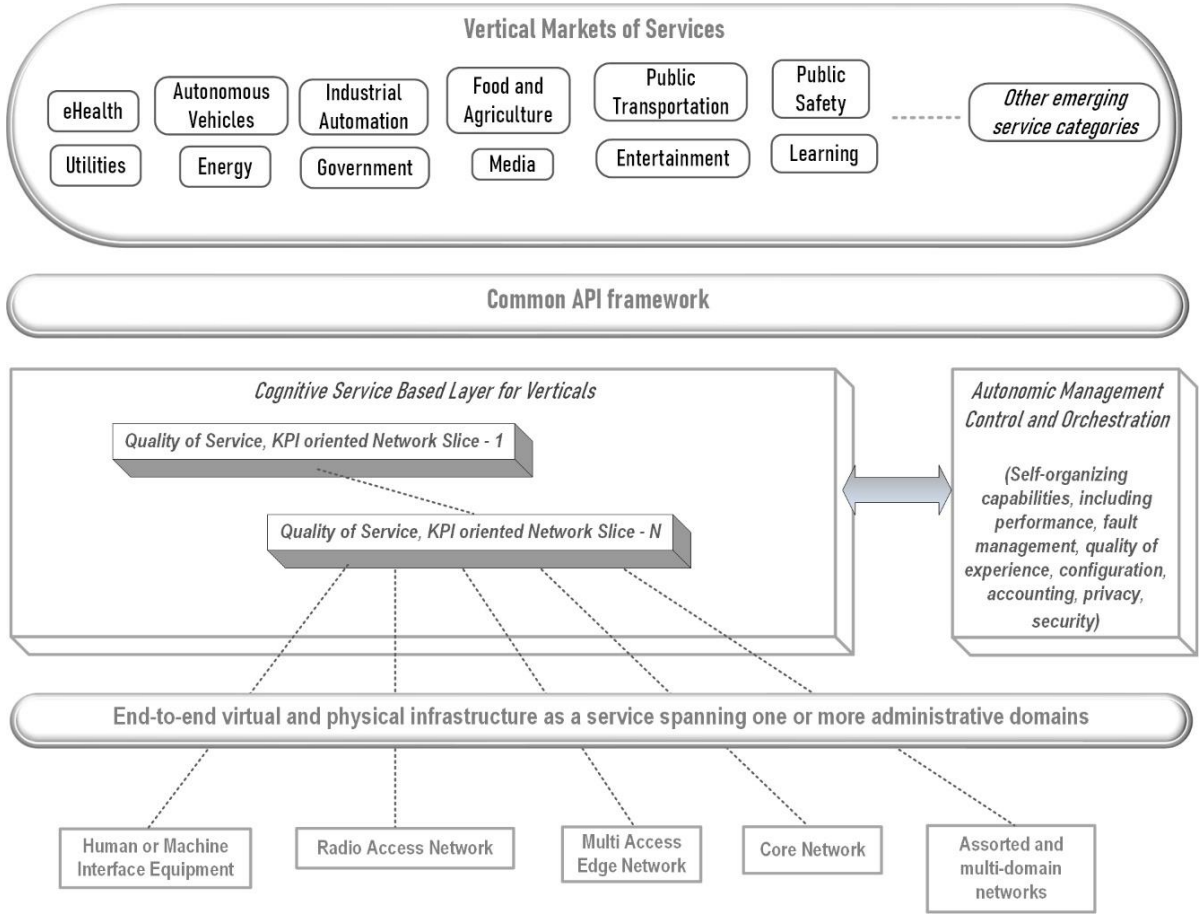


:

Fig. 19 : Ontology of identities and role

10.3 Architectural considerations for a Vertical Market

The enablement of a Vertical market requires a cooperative interaction among the various enabling capabilities that collectively choreograph the operation of end-to-end infrastructure as service. A Vertical market is an ecosystem of diverse services offered by different and emerging categories of Verticals that provide services associated with a given category of services (e.g. Smart City, eHealth etc.). The dominant enabling capabilities are NFV within a service based framework, and the extension of the essence of virtualization through MEC. These dominant enabling capabilities, in concert with autonomic networking provide a foundation for a programmable, software-oriented transformation of the network infrastructure towards a flexible and cognitive service for an expansive and innovative Vertical market of services as illustrated in Fig. 20.



:

Fig. 20 : Cognitive service framework for the Vertical market

The MEC enabler provides a virtualized platform for the execution of services at a distributed network edge to augment the service quality and experience for human and machine interfaces, respectively. The NFV realization through a service based framework provides a platform for virtual network functions. The service based framework in concert with MEC provides a customizable network slicing environment for supporting the demands of diverse service offerings associated with the Vertical market. The end-to-end infrastructure as a service is expected to serve as a significant catalyst in acceleration of the Vertical market engagement and participation, along an increasing value-added trajectory for providers of services, end-users, and for services over a plethora of sensors and actuators in an evolutionary IoT landscape. The network slice serves as a dominant virtualized resource allocation scheme for realizing a logical end-to-end network, which can be

dynamically instantiated to suit the demands of a service in the Vertical market (e.g. stringent QoS associated with an URLLC type of service, security/privacy aspects of a private network etc.). The network slice [3] [4] composition consists of the required instances of network functions, network resources, spectrum resources.

For a flexible integration with diverse service offerings in the Vertical market, a common API framework is a significant component of an end-to-end infrastructure as a service. It is anticipated that as the service offerings the Vertical market will continue to evolve and emerge, the capabilities of the common API framework require to be appropriately augmented. Beyond the rising demand for higher bandwidths, improved coverage, lower latencies, and advanced predictability, reliability, and availability for connectivity, innovative and emerging services require adaptable levels of customization for an enhanced service experience and quality for human and machine interfaces. For example, emerging usage scenarios, such as Augmented Reality (AR) and Virtual Reality (VR) implicitly require stringent KPIs to be satisfied for a sustenance of experiential service quality. This demands an adaptable leveraging of the networking, computing, storage resources that are distributed at the network edge.

The common API layer in the cognitive service framework shown in Fig. 20 is a part of the end-to-end system for initiating the required functionality through service exposure, for accessing the services and data required to satisfy the resources associated with the launching of a service in a Vertical domain. The security of the underlying system and the privacy aspects are preserved through secure information hiding schemes, in a service exposure environment.

The cognitive service framework supports the QoS and the QoE requirements of a service in a Vertical domain, through the configuration, management, control, and instantiation of end-to-end network slice operation, which involves a cooperative resource allocation and management within a given NSP or across collaborating NSPs. The network slice operation, based on a service based architecture, combined with a cognitive service layer, allows for a high granularity of flexibility for Verticals to efficiently and easily create and/or upgrade human and/or machine interface oriented services.

The cognitive awareness is embedded within the virtual network functions of the service based framework, which collaborate and cooperate with autonomic management and control, through fast and slow feedback loop adapted to suit the required resource allocation for supporting the KPIs associated with the realization of a service offering provided by a Vertical domain (e.g. eHealth, industrial automation, NPN (Non Public Network), NTN (Non Terrestrial Network, Smart Home etc.). The virtual functions of the service based framework, utilize SDN principles, together with service based interactions among control plane functions for a realization of cloud-native enabling behaviors. The service based framework designed with the enablement of a microservice [48] approach as needed, provides a high-level of flexibility and granularity for supporting the appropriate network slice composition for Verticals. Microservices facilitate these attributes of flexibility and granularity through malleable levels of functional composition and decomposition as smaller, independent service components, which allow for a rapid implementation of system objectives and system integration.

This microservice enablement allows for the embedding of cognitive awareness for the appropriate levels of infrastructure as a service support for a given, forward-looking service in a Vertical market. High levels of agility and scalability, amenable for rapid and convenient adoption, are established through a decomposition of services into a mesh of component services with an associated smaller set of functional objectives (i.e. Cloud Native Functions (CNFs), where each component is encapsulated as containers within a microservice, which obviates the need for multiple virtual machine instances) Besides the benefit of flexibility and granularity, the use of containers also promotes the customization of VNFs as needed. A microservice may be instantiated multiple times within an end-to-end system, without requiring the traditional cumbersome integrative procedures, promoting a rapid time-to-market of service offerings in the Vertical market.

11 END-TO-END SECURITY

11.1 General

This section provides an overview of security consideration for the protection of the various features and enabling capabilities described in this document, from an end-to-end perspective in a forward-looking 5G service paradigm.

The different components of security in a 5G system, include network access security, network domain security, user domain security, application domain security, visibility, and configurability of security. Furthermore, in the 5G system, the cryptographic algorithms of a 4G system are reused, with the same key length of 128-bit for the protection of the control plane, user plane and RRC signalling. Studies are in-progress to extend the key length for 256-bit protection over the air interface in the future.

11.2 Autonomic networking

The various features and capabilities in the end-to-end system has motivated the introduction of new concepts and protocols, associated with a service-based architecture, which hinges on functional virtualization. For example, network functions in a service-based architecture are interconnected using technologies that are inherent in web services.

The Service Based Interfaces (SBIs) are protected either with Transport Layer Security (TLS) and should support server-side and client-side certificates or could rely on Network Domain Security (NDS), which could be used also in addition to TLS. Furthermore, the interconnect security may also be realized with web-based technologies using JSON Web Encryption (JWE) so that IPX providers are still able to modify the cleartext part of the HTTP message and apply JSON Patch objects.

The security impacts of virtualisation are part of ongoing studies study [54], including the Security Assurance Methodology (SECAM) and Security Assurance Specification (SCAS) for virtualized network products [55].

Since the context of an autonomic framework, which has inherent cognitive capabilities for a realization of system-wide self-CHOP behaviours, depends on network slicing, the related network slicing security consideration is significant. Network slice authentication can be performed, in addition to the primary authentication, which is based on the requested S-NSSAI and slice subscription data, where a third-party application provider can authenticate the access to its slice.

Slice authentication is performed after primary authentication at the time of registration for access to the network. Here the AMF acts as the authenticator and the AAA server of the service provider acts as an EAP authentication server. The authentication signalling between UE and AAA server is routed based on the S-NSSAI via the AMF and the Network Slice Specific Authentication and Authorization Function (NSSAAF). The NSSAAF undertakes any AAA protocol interworking with the AAA server. Multiple EAP methods are possible for slice specific authentication. If the AAA server belongs to a third party the NSSAAF interacts with the AAA server, via a AAA-Proxy (AAA-P). The NSSAAF and the AAA-P maybe co-located.

11.3 Network Data Layer

The shift towards a cloud-native system, hinging on functional virtualization, facilitates a programmable environment for flexible deployment arrangements of decentralization and distribution. In such as system, the operational efficiency of the system demands a unified and cloud-oriented handling of data securely in the system for the network data layer.

The primary authentication enables the mutual authentication between the UE and the network to allow access for any UE (5G capable UE) to NSP rendered services. The 5G system supports EAP-AKA' and 5G AKA as mandatory authentication methods. For a 5G private network additional authentication method, for

example EAP-TLS or any EAP method may also be used. It is the decision of a serving NSP to select an authentication method for the primary authentication. The main features and purpose of the primary authentication are listed as follows:

- Serving Network Authorization
  AUSF gets the serving network name from the SEAF during primary authentication. Before using the serving network name, AUSF checks whether the SEAF is authorized to use it by comparing the serving network name with the expected serving network name to ensure the authenticity of the serving network.

- Serving Network Binding
  The serving network specific anchor key (K_{SEAF}) is bound to the serving network by including the serving network identifier in the key generation to prevent the serving network from claiming to be a different serving network (e.g. fraudulent charging). Hence this approach provides an implicit serving network authentication for the UE. The binding of a specific anchor key (K_{SEAF}) to the serving network provides assurance for the UE that it is indeed connected to a serving network, which is authorized by the home network (e.g. home NSP of the UE).

- Robust home network control
  The authentication and key agreement protocols for a service-based system provide a robust home control scheme, since every authentication result is reported to the Unified Data Management (UDM), which is part of the network data layer. Robust home network control is useful for the prevention of certain types of fraud (e.g. fraudulent N_{UDM} - UECM Registration Request for registering the subscriber's serving AMF in the UDM that is not actually present in the visited network). The actions taken by the home network to link the authentication confirmation or the absence of a confirmation using subsequent procedures, which are not standardized, are subject to the home network policy.

- Multiple NAS Connection
  A primary authentication over a trusted access network can also provide keys to establish a secure association between the UE and an N3IWF, which is used for access to an untrusted access network

The study on authentication enhancements is ongoing [56], where the focus is on the following prominent aspects:

- Mitigation of linkability attack
- Fraud attack resulting from an expired authentication in the UDM
- SUPI guessing attack
- Protection of Sequence Number (SQN) during AKA re-synchronization

Recently discovered attacks in 4G systems, were found to exploit the absence of integrity protection in the UP, are also a relevant vulnerability for 5G systems. In 5G (Phase-1), there were only two code points that were standardized for integrity protection of the UP, namely, a mandatory minimum support of 64 Kbps or full rate support. In 5G (Phase-2), all UEs are required to support a full rate integrity protection of the UP.

The usage of integrity protection hinges on the NSP policy and is signalled to the gNB and the UE, within the associated security policy. The SMF provides the UP security policy for a PDU session to the ng-eNB/gNB, during the PDU session establishment procedure. The UP security policy indicates whether a UP confidentiality and/or UP integrity protection have to be activated or not for all DRBs belonging to that PDU session. If the UP security policy indicates "Required" or "Not needed", the ng-eNB/gNB should not overrule the UP security policy provided by the SMF. If the ng-eNB/gNB cannot activate UP confidentiality and/or UP

integrity protection when the received UP security policy is "Required", then the gNB will reject establishment of UP resources for the PDU Session and indicate reject-cause to the SMF. The UP security is activated at the UE by the ng-eNB/gNB with a RRC Connection Reconfiguration procedure.

For network operation security, it should include the security control of physical equipment and the security control of business data. It should have a complete security policy based control system, which can fulfil holistic and multi-level monitoring and protection requirement through the access control system, signalling monitoring system, firewall/IDS/IPS protection, etc. With these mechanisms, the operation of the end-to-end system should be able to ensure the reliability, security, efficiency, and traceability of event behaviour of data transmission and storage at the network data layer.

11.4 AI and ML

The adoption of AI and ML for system-wide cognitive capabilities for a realization of self-CHOP behaviours in autonomic networking for an orchestration, optimization, and automation of services. On-demand network slicing is leveraged to satisfy user service requirements.

The NWDAF collects data and analytics from various cognitive functions and the UE, and for a further processing of analytics, using AI and ML capabilities, to infer and deliver appropriate analytics to the subscribed network functions based on the corresponding analytics subscription and request. The collection of UE data to fulfill NWDAF functionalities may breach the user information security, such as privacy, integrity and accessibility aspects and violation of user consent.

As part of autonomic networking, cognitive capabilities in the system are anticipated to evolve in a distributed manner, where these cognitive capabilities are imbued in both physical and virtual functions. Further studies are ongoing to address the security concerns associated with the collection of data and its exposure to consumers of analytics information, data collection and exposure of data to the analytics consumers. Enhancement of NWDAF functionalities can be leveraged to support the prediction and detection of cyber-attacks and anomalies in the network behaviour, operation and function.

11.5 Virtualization in the RAN

The virtualization of the RAN, such as BBU resources, there are related vulnerabilities that require to addressed. A review of virtualized RAN security risks and potential solutions is examined in [40]. Relative to a centralized and integrated traditional RAN architecture, a virtualized RAN architecture of BBUs and RRUs, hinges on a decoupling of software and hardware for flexible deployment choices and ease of configurability and system integration, in terms of changing functional split arrangements to suit corresponding changes in market and service demands.

On the other hand, in the case of a distributed arrangement of BBUs and RRUs, trust is not implicit, especially if the virtualized RAN is shared by multiple NSPs or SPs. While end-to-end network slicing capabilities implicitly isolate resources, methods of trust establishment are necessary, where multiple domains need access to any shared resource. Automated methods of establishing trust through the use of suitable types of a distributed ledger and smart contracts may provide potential solutions for the establishment of trust in cases where a virtualized RAN is shared by multiple domains [31].

Integrated Access and Backhaul (IAB) and disaggregated gNB are prominent features in the RAN. In the case of IAB, an IAB-node acting as an IAB-UE, wirelessly connects to a gNB, referred to as an IAB-donor, which is capable of serving IAB-nodes. IAB utilizes the CU/DU split architecture, where the IAB-node terminates the DU functionality, and the IAB-donor terminates the CU functionality. Essentially, the IAB-node is a relay for signalling and traffic between the UE and the RAN, and represents a gNB towards the UE, and as an IAB-UE towards the IAB-donor in the RAN.

The RAN is a network of access and backhaul links as specified [57].The authorization of the IAB-nodes performed by the core network, where IAB is supported. The IAB-node, acting as UE, registers with the network to establish both NAS and AS security. For securing the F1 interface between the CU and the DU, IKEv2 PSK authentication may be utilized, and to further support a plug-and-play configuration of an IAB-node and an IAB-donor, without any pre-configuration, dynamic PSK may also be supported. The support for DTLS for securing the F1 interface is optional for the IAB-node and he IAB-donor [58].

In the case of disaggregated gNB, a gNB may consist of a gNB-CU-CP, multiple gNB-CU-UPs and multiple gNB-DUs. The gNB-CU-CP selects the appropriate gNB-CU-UP(s) to support the requested services for the UE as specified [59].

As currently specified, if any UE has user plane connections established, via more than one gNB-CU-UP, then all user plane data applies the same UP security keys. In the case of Verticals, if any service related data associated with a UE, which is handled by one gNB-CU-UP needs privacy from its other service data that is handled by a different gNB-CU-UP, potentially owned by a different service provider, then the current model may not guarantee the desired level of privacy.

Additionally, if any gNB-CU-UP is compromised then it will breach the security of all gNB-CU-UP's and the user plane security of that UE. To realize a robust system-wide security by design, a gNB-CU-CP based on its local policy can derive and apply a distinct user plane security context, with cryptographic separation across different user plane contexts, to enable a robust user plane security in the system.

11.6 Multi-access Edge Computing

The distribution of networking, computing, and storage resources at the network edge, facilitates the support for a wide variety of services with diverse KPI requirements (e.g. low latencies, high reliability, bandwidth efficiency etc.), over converged access (e.g. terrestrial and non-terrestrial). Studies are ongoing in terms of threats, security requirements and solutions, where the MEC platform is associated with the UPF. The Edge Enabler Server (EES) provides functionalities for the Edge Enabler Client (EEC) over the EDGE-1 reference point, such as configuration information for the EEC, as well as support the functionalities of application context transfer.

The EEC retrieves functionalities, such as configuration information, from the EES, together with discovering edge application servers that are available in the Edge Data Network (EDN), which is localized data network. The EDN contains both the EES and EAS(s). The Edge Data Network Configuration server (EDNCS) contains the necessary functions and the configuration information for the EEC to connect with the EES. The EES performs registration, registration update, or de-registration of EES information with the EDNCS. An EES that is configured with multiple EDNCS endpoint addresses may perform the service registration, update, or deregistration procedures in accordance with the EDNCS, where the security context of each of EDGE-6 interfaces needs to be isolated from each other, since the corresponding trust domains may be different.

With respect to the EDN, it is localized and closer to the UE, in terms of the end-to-end system. Depending on the trust relationship between the EDN and the related NSP domain (i.e. PLMN), the authentication of the EDN many be implicit. The ongoing studies [60], focus on enabling the following security aspects to support MEC security in the system. Some of the main aspects of these studies include:

- Mutual authentication between EEC and EES.
- Authorization of EEC by EES to allow access to EES provided services.
- Mutual authentication between EEC and EDNCS.
- Authorization of EEC by EDNCS to allow access to ECS provided services.

- ❖ Mutual authentication between EES and EDNCS to register and update the server profile information.
- ❖ Authorization of EES by EDNCS to allow access to ECS provided services.
- ❖ Mutual authentication between UE and EDN
- ❖ Authorization of UE to access EDN

11.7 Distributed Ledger Technology

Periodic security checks and/or verification is necessary with respect to routing tables and the configuration associated with the DLT platform. The distributed ledge nodes are to be deployed, such that an impersonation of the participating nodes is prevented. The transactions at any given location in the distributed ledger is subject to verification and a limiting transaction volume threshold.

Data categorization and checking is necessary before admittance to a distributed ledger. Sensitive data should not be uploaded into a distributed ledger or should least be ciphered before an introduction into the distributed ledger. To prevent the pollution of a distributed ledger, spam data should not be uploaded into the distributed ledger. Robust blockchain protocols should be applied to the distributed ledger to mitigate the impact of security threats, such as:

- ❖ Timestamp checking to avoid a subsequent tampering of a distributed ledger record
- ❖ Mining reports to reveal the integrity of the distributed ledger and to mitigate attempts to corrupt the distributed ledger
- ❖ Robust hashing algorithms and cryptographic algorithms to prevent hash function collision and cracking of an existing block hash in the distributed ledger

11.8 Vertical Market

The Vertical market is an expansive ecosystem of emerging services with diverse KPI demands, for a variety of industries, over a virtualized and cognitive end-to-end system that facilitates performance and service experience optimization. The security considerations to support the Vertical market, includes the following:

- ❖ Secondary Authentication

The secondary authentication is performed after a successful primary authentication with an external Data Network (DN), which is located outside the NSP domain. The procedure is related to the establishment of the UP associated with the external DN, and is based on the EAP framework with pre-provisioned credentials from the DN. The SMF takes the role of the authenticator and the AAA Server in the DN takes the role of the EAP authentication server. Once the secondary authentication is successful for a given UE, the UP connection is established, and the service of the external DN can be supported.

- ❖ AKMA Security

Authentication and Key Management for Applications (AKMA) enables the extra level of authentication, which targets Internet of Things (IoT) devices that have restricted resources and reduced capabilities. AKMA enables authentication and key management aspects for applications based on subscription credential(s) in the system.

Authentication frameworks, such as GBA [61] and BEST [62] applicable in pre-5G systems, are not supportive of the service-based architecture, virtual network functions, and the emerging IoT application layer protocols.

- ❖ The AKMA-specific authentication is performed after a successful primary authentication and the AKMA key is derived from the AUSF key. Furthermore, when required, an application function

specific key (AKMA application key) can be derived from AKMA key: *NPN Authentication and Key storing Security.*

Additional requirements from a security architecture perspective to enable support for a Standalone NPN (SNPN) along with subscription/credentials owned by an entity separate from the SNPN, may require changes in the primary authentication procedure in the 5G System.

Studies are ongoing with respect to credential management to support UE onboarding and provisioning for an NPN

- Deterministic/Timing Synchronization Network Security
  As part of emerging Verticals, the security for Time Sensitive Communications (TSC), in a Time sensitive Network (TSN) is pivotal, for the enablement of deterministic and low-latency communications to support diverse services stringent KPIs that are anticipated, for example, in Industry 4.0 [39] usage scenarios and factory automation

  In broad terms, TSC is among a variety of MEC usage scenarios, especially those that have the characteristics of the URLLC service category. A TSC enabled UE should communicate with a TSC bridge or another 5G TSC enabled UE in a protected way. The messages from the Device Side - TSN Translator (DS-TT) in the UE to the Network Side – TSN Translator (NS-TT) in the UPF are secured with UP security, between the UE and the gNB. The corresponding UP security enforcement information should be set to 'required'.

11.9 Privacy

The Subscription Unique Permanent Identifier (SUPI) of the UE is privacy protected to enable the privacy of the subscriber and is never sent as clear text over the air, during the primary authentication process to prevent IMSI catching and user tracking. The primary authentication process supports an exchange of the Subscription Concealed Identifier (SUCI) between the UE to the network, where the UE is identified by extracting the SUPI from the SUCI, using Subscription Identifier De-concealing Function (SIDF) based service, offered by the Unified Data Management (UDM) in the home network.

In the case of the IP Multimedia Subsystem (IMS), the IP Multimedia Private Identity (IMPI) is a unique permanently allocated global identity assigned by the home network NSP. The IMPI has the form of an Network Access Identifier(NAI) i.e. user.name@domain. The username typically contains the SUPI. When IMS traffic transverses the network [4], subscriber privacy is assured by the underlying air interface and network domain security features. Precautions need to be taken when the IMS traffic transverses other networks and certain identity information of the subscriber is allowed to be withheld as specified [63] [64]. Other aspects of privacy, such as ensuring Informed User Consent for services are for future study. Other studies, include:

- False Base Station (optional)

The study on 5G security enhancements for protection from false base stations [65] is ongoing for different ways to detect false base stations in the network and in the UE. So far, a network approach is based on measurement reports that the UE is sends in the RRC Connected mode.

- Usage scenarios
  - Security of Cellular Internet of Thing (CIoT)
    The evolution of security for CIoT includes enhancements [66], such as:
    - ✓ Security handling in Control Plane CIoT 5G system optimization is performed in the NAS connection reusing NAS security context for integrity protection and ciphering for the small user data or SMS between the UE and the AMF.

- ✓ For the protection of the RRC re-establishment procedure, the AS part of the UE triggers the NAS part of the UE to generate the UL_NAS_MAC and XDL_NAS_MAC of the RRC message as if it would be a NAS message.

- Security of Vehicle to Everything (V2X)
  The security aspects of advanced V2X services [67] are limited to communications in the unicast mode. For unicast, the UE is preconfigured with long term credentials in order to authenticate with other UEs and to derive the key material for the protection of the direct communication. The unicast mode communication for the signalling and user plane can be protected for integrity and confidentiality at the PDCP layer.

There are no requirements and procedures for securing the NR based PC5 reference point for groupcast and broadcast mode. For avoiding link ability and trackability attacks, the UE should only change and randomize its source Layer-2 ID and source IP address, including IP prefix (if used).

12 LIST OF ABBREVIATIONS

3GPP	Third Generation Partnership Project
4G	Fourth Generation 3GPP system
5G-RG	5G Residential Gateway
AAA	Authentication, Authorisation and Accounting
AAA-P	AAA-Proxy
AD	Autonomic Domain
AI	Artificial Intelligence
AI-SS	AI-Support System
AKA	Authentication and Key Agreement
AKMA	Authentication and Key Management for Applications
AM	Autonomic Manager element, which is also referred to as a DE
AMF	Access and Mobility Function
API	Application Programming Interface
AUSF	Authentication Server Function
BBU	Base Band Unit
BSS	Business Support System
CBRS	Citizens Broadband Radio Service
cDE	Centralized DE
CI/CD	Continuous Integration/Continuous Delivery
CN	Core Network
CP	Control Plane
CPE	Customer Premises Equipment
CPS	Control Plane Service
CSP	Any provider of communication services, and includes SP and NSP
CU	Centralised Unit
D2D	Device-To-Device
dDE	Distributed DE
DDoS	Distributed Denial of Service
DE	Decision-making Element, which is also referred to as an AM element
DID	Domain Identifier
DN	Data Network
DoS	Denial of service
DPL	Deep Learning
DT	Domain Type

DU	Distributed Unit
E2E	End-to-End
EAP	Extensible Authentication Protocol
EAS	Edge Application Server
EDGE-1	Reference point between the EES and the EEC
EDGE-6	Reference point between the EDNCS and the EES
EDN	Edge Data Network
EDNCS	EDN Configuration Server
EEC	Edge Enabler Client
EES	Edge Enabler Server
EM	Element Manager
eMBB	Enhanced Mobile Broadband
ETSI	European Telecommunications Standards Institute
FCAPS	Fault, Configuration, Alarm, Performance and Security Management.
FFT	Fast Fourier Transform
FMC	Fixed Mobile Convergence
FTTH	Fibre To The Home
GANA	Generic Autonomic Networking Architecture
gNB	gNodeB (5G base station)
GSM	Global System for Mobile Communications
GSMA	GSM Association
GUTI	Globally Unique Temporary Identifier
H-H	Human to Human
H-M	Human to Machine
IaaS	Infrastructure as a Service
IAB	Integrated Access and Backhaul
ICT	Information and Communication Technology
IDS	Intrusion Detection System
IEEE	Institute of Electrical and Electronics Engineers
IFFT	Inverse FFT
IMEI	International Mobile Equipment Identity
IMPI	IP (Internet Protocol) Multimedia Private Identity
IMS	IP (Internet Protocol) Multimedia Subsystem
IMSI	International Mobile Subscriber Identity
IPS	Intrusion Prevention System
IPX	Internetwork Packet Exchange
ITU	International Telecommunication Union
Kafka	Publish/subscribe software message bus for a scalable handling of lifecycle management events associated with service orchestration
KP	Knowledge Plane
KPI	Key Performance Indicator
KQI	Key Quality Indicator
LCM	Life Cycle Management
LTE	Long Term Evolution
MANO	Management and Orchestration
MBTS	Model Based Translation Table, which is an intermediation layer between the KP DEs (Decision Elements) and the NEs (Network Elements - physical or virtual) for translating vendor-specific raw data onto a common data model for use by network level DEs [9]
ME	Managed Entity

MEC	Multi-access Edge Computing
mIoT	massive Internet of Things, typically referring to 5G IoT
ML	Machine Learning
M-M	Machine-to-Machine
MMTEL	Multimedia Telephony
N3IWF	Non-3GPP Inter-Working Function
N6	3GPP interface between the 5G core network and a Packet Data Network
NaaS	Network as a Service
NE	Network Element
NEF	Network Exposure Function
NETCONF	Network Configuration Protocol, which utilizes Yang to model the operations, such as network configuration, state data, Remote Procedure Calls (RPCs), and notifications.
NFV	Network Function Virtualisation is a model for the virtualisation of network functions, realized in software [7]
NFVI	NFV Infrastructure
NFVO	NFV Orchestrator
ng-eNB	next generation - eNB (evolvedNodeB)
NPN	Non-Public Network. 5G private network that provides network services for a given organization or a group of organizations.
NSP	Network Service Provider, is an entity that provides network resources and services
NSSAAF	Network Slice Specific Authentication and Authorization Function
N_{UDM} - UECM	Service-based interface for UE context management
NWDAF	Network Data Analytics Function
OLA	Operations Level Agreement
ONIX	Overlay Network for Information eXchange) useful for enabling auto-discovery of information/resources of an autonomic networking, via “publish/subscribe/query and find” protocols [9]
OSS	Operation System Support
OTT	Over-The-Top
PaaS	Platform as a Service
PC5	Interface for direct communications between vehicles and any other device
PCF	Policy Control Function
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDU	Protocol Data Unit
PDV	Packet Delay Variation
PELR	Packet Error Loss Rate
PNF	Physical Network Function
PON	Passive Optical Network
PUF	Physical Unclonable Function
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
REST-ful API	Representation State Transfer oriented API
RRC	Radio Resource Control
RRU	Remote Radio Unit
RU	Residential Unit
SBA	Service-Based Architecture [4]

SC	Smart Contract
SDN	Software Defined Networking
SDO	Standards Developing Organization
SEAF	Security Anchor Function
SEPP	Security Edge Protection Proxy
SIDF	Subscription Identifier De-concealing Function
SLA	Service Level Agreement
SMF	Session Management Function
S-NSSAI	Single – Network Slice Selection Assistance Information, which uniquely identifies a network slice.
SON	Self-Organising Network
SON-C	SON-Centralized for self-organizing capabilities in backhaul transport for wireless access
SP	Service Provider, is an entity that provides services, and may or may not include a network infrastructure
SUCI	Subscriber Concealed Identifier
SUPI	Subscriber Permanent Identifier
TMSI	Temporary Mobile Subscriber Identity
TNAP	Trusted Non-3GPP Access Point
TNGF	Trusted Non-3GPP Gateway Function
TSC	Time Sensitive Communications
TSN	Time Sensitive Networking
UDM	Unified Data Management
UDR	Unified Data Repository
UDSF	Unstructured Data Storage Function
UE	User Equipment (human or machine interface)
UN SDG	United Nations Sustainable Development Goal
UN SDSN	United Nations Sustainable Development Solutions Network
UP	User Plane
UPF	User Plane Function
UPS	User Plane Service
URLLC	Ultra-Reliable Low Latency Communication
V2X	Vehicle to Everything
VIM	Virtualized Infrastructure Manager
VNF	Virtualised Network Function is a building block implementation of a network function using software that is decoupled from the underlying hardware, such as in a cloud-native environment (e.g. SBA framework that leverages the model of NFV)
VNFM	VNF Manager
W-AGF	Wireline-Access Gateway Function
WIM	WAN (Wide Area Network) Infrastructure Manager
X-Haul	Flexible, heterogeneous access fronthaul and backhaul
YANG	Yet Another Next Generation (Data modelling language for the definition of data sent over network management protocols)