SEMANTIC

end-to-end Slicing and data-drivEn autoMAtion of Next generation cellular neTworks with mobIle edge Clouds

Marie Skłodowska-Curie Actions (MSCA)
Innovative Training Networks (ITN)
H2020-MSCA-ITN-2019
861165 - SEMANTIC

WP3 – Optimizations for integrated access/X-haul and end-to-end slicing

D3.1: SoA on integrated access/X-haul and e2e slicing

Contractual Date of Delivery:	M18
Actual Date of Delivery:	30/06/2021
Responsible Beneficiary:	CTTC
Contributing beneficiaries:	CTTC,EUR,IQU
Security:	Public
Nature:	Report
Version:	V2.1

Document Information

Version Date: 30/06/2021
Total Number of Pages: 38

Authors

Name	Organization	Email
Swastika Roy	CTTC	sroy@cttc.es
Pavlos Doanis	EURECOM	pavlos.doanis@eurecom.fr
Suvidha Sudhakar Mhatre	IQUADRAT Informatica	s.mhatre@iquadrat.com
Dr. Luis Blanco	CTTC	lblanco@cttc.es
Dr. Christos Verikoukis	CTTC	cveri@cttc.es

Document History

Revision	Date	Modification	Contact Person
V1.0	21/05/2021	Submitting the draft version	Ms. Swastika Roy
V1.0.1	23/06/2021	Editor changes	Dr. Luis Blanco
V1.1	28/06/2021	Merging contributions of all ESRs	Ms. Swastika Roy
V2.0	30/06/2021	Editor changes	Dr. Luis Blanco
V2.1	30/06/2021	Merging contributions of all ESRs	Ms. Swastika Roy

Table of Contents

List of Figures

List of Acronyms and Abbreviations

Acronym	*Description*
3GPP	Third Generation Partnership Project
5G	5th generation (5G) mobile network
5G NR	5G New Radio
ADMM	Alternating Direction Method of Multipliers
AI	Artificial Intelligence
BAP	Backhaul Adaptation Protocol
BBU	Base Band Unit
BH	Backhaul
CN	Core Network
CPU	Central Processing Unit
C-RAN	Cloud Radio Access Network
CU	Centralized Unit
CU-CP	Centralized Unit – Control Plane
CU-UP	Centralized Unit – User Plane
CSMF	Communication Service Management Function
DAG	Directed Acyclic Graph
DC	Data Center
DL	Downlink
DU	Distributed unit
E2E	End-to-End
eMBB	Enhanced Mobile Broadband
ETSI	European Telecommunications Standards Institute
F1-C	F1 interface for the Control plane
F1-U	F1 interface for the User plane
gNB	5G base station
IAB	Integrated Access and Backhaul
IAB-DU	IAB-Distributed Unit
IAB-MT	IAB-Mobile Terminal
IBN	Intent-Based Networking
ILP	Integer Linear Program
ITU-T	International Telecommunication Unit- Technical Standardization
IMT	International Mobile Telecommunications
MDs	Management Domains
mMTC	Massive Machine Type Communication
MILP	Mixed Integer Linear Program
ML	Machine Learning
mmWave	millimeter Wave
NG-RAN	New Generation – Radio Access Network
NSMF	Network Slice Management Function
NSSMF	Network Slice subnet Management Function
NSI	Network slice instance
NSSI	Network slice subnet instance
PDCP	Packet Data Convergence Protocol
PtP	Point to Point

RAN	Radio Access Network
RL	Reinforcement Learning
RLC	Radio Link Control
RRC	Radio resource Control
RRH	Remote Radio Head
SDN	Software-Defined Networking
SFC	Service Function Chain
SLA	Service Level Agreement
TN	Transport Network
TSG SA	Technical specification group service and system aspects
UE	User Equipment
UL	Uplink
URLLC	Ultra-reliable low latency communication
V2X	Vehicular to everything communication
VNE	Virtual Network Embedding
VNF	Virtual Network Function
VNF-FG	VNF Forwarding Graph
VNF-FGE	VNF-FG Embedding
QoS	Quality of Service
ZSM	Zero Touch Network and Service Management

1 Executive summary

This deliverable summarizes the objectives of WP3 (Optimizations for integrated access/X-haul and end-to-end slicing) contributed by the SEMANTIC ESRs. More precisely, it outlines the major outcomes of the ESRs towards Task 3.1 (SoA on integrated access/X-haul and e2e slicing) that includes the concept of network slicing, IAB, state of the art of end-to-end network slicing, as well as current standardization efforts on network slice management and orchestration fields. At the same time, it indicates some open research areas and challenges related to the topics mentioned above.

2 Introduction

2.1 End-to-end slicing

Upcoming wireless networks demand seamless high connectivity, data speed, quality, ultra-low latency and reliability which will connect ¨Everything¨ (human, machines, cloud, server) referred to as Internet of Everything. 5G advance feature are blurring the line between traditional user association hence require better load balancing going beyond coverage area of specific cell. The 5G and beyond network must be able to offer capabilities for diverse requirements such as ultra-reliable services, ultra-high-bandwidth, extremely low latency at the same time. Slicing can achieve the most efficient utilization of available resources in 5G and beyond networks. It permits customers/end users to enjoy connectivity and data processing tailored to the specific vertical requirements that adhere to a Service Level Agreement (SLA) agreed with the mobile operator. End-to-end (E2E) network slicing is one of the enabling technologies to achieve these various requirements of upcoming beyond 5G and 6G networks. E2E slicing can be defined as an independent end-to-end logical network that runs on a shared physical infrastructure and capable of providing a negotiated service quality. Network slicing is key enabler for the next generation of wireless networks. It allows:

- Capability to provide diverse requirement at the same time and simultaneous services to different verticals.
- Efficient utilization of available resources in the network.
- Connectivity and data processing tailored to an SLA.
- Easy customization of network as per service and required capability.
- 5G advanced features, such as multi connectivity, virtual cells require better deployment of load balancing in terms of distribution of resources and traffic among the network elements.

Slicing provides network connection service offered to business customers at a connectivity level such as near real time latency, reliability, etc. In addition to this slicing can enable network resource services that grants access to the operator network resources for running proprietary applications like cloud computing, positioning, edge computing, security, etc.

2.1.1 Architecture (as per 3GPP)

5G network slicing concept enables the creation of multiple unique logical and virtualized networks over a shared physical infrastructure in an efficient and economical way. Its main purpose is to provide differentiated network services, by providing isolated end-to-end virtual networks tailored to specific requirements (e.g. data speed, quality, latency, reliability) and SLAs that serve users with very different needs.

A network slice could span across multiple network domains (e.g., RAN, edge, cloud and transport network) and could also be deployed across multiple operators. A network slice comprises dedicated and/or shared resources in terms of processing power, storage, and bandwidth.

Slice types could be defined from a functional or behavioural perspective. Mobile network operators could deploy a single network slice type that satisfies the needs of multiple verticals, as well as multiple network slices of different types with multiple and diverse requirements (for example a vehicle may need simultaneously a high bandwidth slice for infotainment and an ultra-reliable slice for telemetry, assisted driving).

The 5G technology considers three generic services with vastly heterogeneous requirements:

1) Enhanced mobile broadband (eMBB) -high data rates requirements
2) Massive machine-type communications (mMTC) -need to support a very large number of devices in a small area
3) Ultra-reliable low-latency communications (URLLC) -strict requirements on latency and reliability

Consequently, service heterogeneity can be accommodated by network slicing, through which resources are allocated to each service to provide performance guarantees and isolation from the other services.

The main challenges referring to end-to-end slicing could be summarized to:

- Automation of slice deployment - operation and maintenance
- Cross-domain connection
- SLAs guarantees that defines the level of service expected from a vendor, measured with specific metrics by which service is measured, as well as remedies or penalties should agree-on service levels not be achieved.

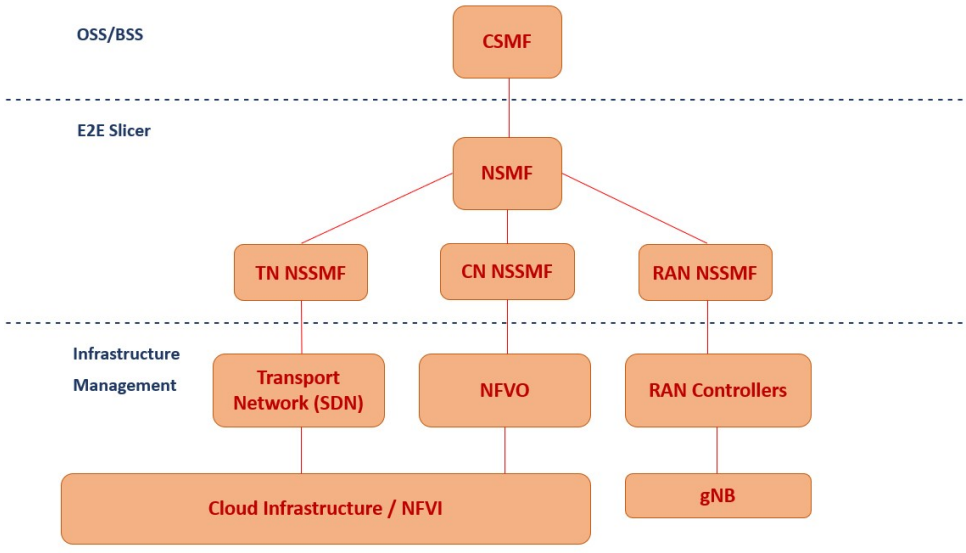

Figure 2-1: E2E Slicing Architecture (3GPP) [1]

An end-to-end slice should be orchestrated and managed. Introduced Figure 2-1shows the E2E network slicing architecture proposed by 3GPP.The 3GPP has identified 3 different management functions related to network slicing management, which are defined as follows:

- Communication Service Management Function (CSMF): responsible for translating the communication service-related requirement to network slice related requirements.
- Network Slice Management Function (NSMF): responsible for the management and orchestration of NSIs. Moreover, it derives network slice subnet related requirements from the network slice related requirements. The NSMF communicates with the NSSMF and the CSMF.
- Network Slice Subnet Management Function (NSSMF): responsible for the management and orchestration of NSSIs. The NSSMF communicates with the NSMF.

The CSMF is deployed in the OSS/BSS. Upon receiving a request for a communication service and translating it into the respective slice requirements, it interacts with an End-to-End (E2E) Slicer.

The E2E Slicer has a dominant role in network slice selection while also is responsible to collect runtime monitoring information in order to perform slice lifecycle management operations. Tools from artificial intelligence and data analytics could be used to this end.

The interaction between CSMF and E2E Slicer is achieved via NSMF, through a northbound interface, in order to request the instantiation of an end-to-end network slice with specific characteristics. The NSMF then proceeds to the respective slice creation by assigning the instantiation and management of each sub-slice to the appropriate NSSMF. Consequently, the NSMF based on the requested slice type, selects the appropriate NSSMF that correspond to each sub-slice. Each slice template includes fields that indicate what type of NSSMF should be used to orchestrate the underlying sub-slice instance.

2.1.2 Types of subnets and management entities for each subnet

The creation of a network slice is referred as Network Slice Instance (NSI). Each NSI carry the business requirements of end-to-end slicing. In the context of end-to-end network slicing in 5G, a slice will consist of components, in the form of Virtual Network Functions (VNFs), from the Radio Access Network (RAN), the Core Network (CN), and the Transport Network (TN).

A slice is composed by a set of sub-slices, named Network Slice Subnet Instances (NSSIs). A sub-slice is composed by a set of virtual or physical resources from the same technological domain. Regardless of the underlying system and the functional requirements, an NSI should always be composed of three sub-slices: RAN NSSI, CN NSSI and the TN NSSI.

RAN slice-ability gives the opportunity to both its service and resources to be sliceable. The major challenge is the provision of the necessary level of performance isolation across slices, which are sharing typically the same access network and where each slice has its own resource requirements. RAN sub-slices isolation could be achieved through the protocol layers that manage

scheduling and physical radio resource allocation. At the RAN level, the service is radio access and the resources are virtual resource blocks.

The CN sub-slice is responsible for elements related to the CN. These elements could be implemented either in a virtualized way, or through Physical Network Functions (PNFs). The CN virtualization deployment via the NFV Infrastructure (NFVI) gives the opportunity in terms of flexibility to customize the compute and other resources allocated to the slice. Moreover, the slice could scale on demand for cost and performance optimization, and to select the appropriate functional configuration of the CN components (e.g. number of VNF instances for each CN function to maximize reliability). All the aforementioned are supported by maturing relevant standards, such as the ETSI NFV Management and Orchestration (NFV-MANO). The core sub-slice provides EPCaaS, which is supported by specific dedicated virtual storage and compute resources.

Finally, TN sub-slice consists of the transport paths whose management is executed by using technologies such as SDN (Software-Defined Networking) or Virtual LANs. The TN sub-slice provides a connectivity service to external networks, with dedicated virtual network links.

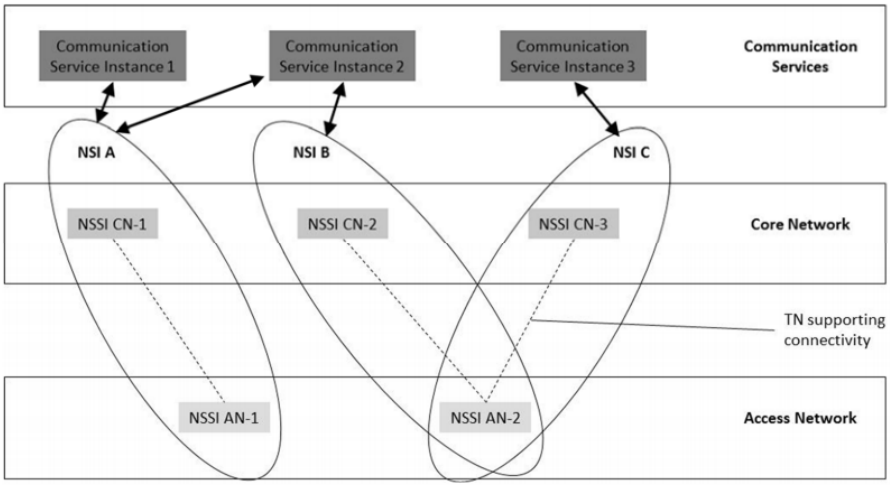

Figure 2-2:A variety of communication services instances provided by multiple NSIs [2]

In case of the core sub-slice, the CN NSSMF, will communicate with the NFVO via a northbound interface in order to request the creation of the appropriate CN NSSI, over the respective cloud/NFVI platform. The interaction with the NFVO includes the launch and configuration of the virtual CN instance (virtual Evolved Packet Core - vEPC network, customization of VNF instances, and allocated resources). The instantiation of all needed VNFs for the creation of the vEPC will be achieved by communicating with the underlying VIM.

In case of the RAN sub-slice, the RAN NSSMF encapsulates a RAN resource allocator which is responsible to translate slice requirements into radio resource allocations and to execute the RAN resource management. In order to accomplish this process, it retains information about RAN's state

(connected UEs per gNB, UE/slice bitrate requirements, radio connection quality, slice instances that a gNB participates). Through this information, in an iterative dynamic way based on RAN state, is determined the appropriate RAN resource partition per cell for the coexisting slices. The communication between RAN NSSMF and gNB is achieved through a north bound interface. The RAN controller can be seen as a RAN-specific Virtual Infrastructure Manager (VIM), while an agent installed at a gNB can be likened as a hypervisor. The E2E slicer is aware of which gNBs correspond to each service area, so that it can appropriately configure the deployed RAN sub-slices.

Finally, in case of the transport sub-slice, the TN NSSMF, is responsible to interact with network elements such as SDN controllers in order to handle the provisioning and isolation of the links connecting (virtual or physical) network functions of the access and core networks, and towards external networks.

2.2 Integrated access and backhaul

The use of millimeter Wave (mmWave) frequencies in 5G wireless networks is one of the key technologies enabling the provision of higher data-rates to the end-users and a larger number of connected UEs. However, the wave propagation characteristics in that part of the spectrum dictate densification of the network due to limited coverage capabilities. Also, more resource blocks per unit area can be provided with a denser deployment of large and small-cell base stations [3], [4].

Traditionally, microwave and optical fiber have been the two backhaul technologies used in the transport network to connect a base station with the Core. Microwave technology is low cost and easy to deploy, but it requires line-of-sight communications due to the high path loss, and supports lower data-rates compared to the fiber.On the other hand, fiber backhaul provides very high capacity but it is difficult and costly to deploy. Historically, the technology that has been mostly used for backhaul is microwave (usually operating in the spectrum above 10GHz) [4].

Network densification in 5G requires the use of radio sites even at the street level. Since fiber backhaul is both very expensive and difficult (or even impossible) to deploy for a very large number of base stations, Integrated Access and Backhaul (IAB) has been employed in 5G networks in order to offer a complementary backhaul technology [3], [4]. In nodes that support IAB, part of the radio resources is utilized for access, while another part of them is utilized for backhaul. Moreover, IAB nodes operate according to the 5G New Radio (NR) air interface and hence they can provide both access and backhaul, even in no line-of-sight scenarios, through the International Mobile Telecommunications (IMT) bands. The most important features of this technology are flexibility, scalability (multi-hop backhaul), automation and fast deployment. 3GPP standardized IAB in Rel-16 while additional standardization activities are currently under discussion for Rel-17. In the following subsections, we are going to outline the architecture of IAB according to Rel-16 and highlight some of its most important features.

2.2.1 Functional splits in C-RAN

Functional Split is one of the new concepts introduced in 5G wireless networks. It facilitates flexibility in base station function placement, with the objective to limit capacity demand on the fronthaul link and support services with strict latency requirements. It is worth making a brief introduction to functional split since IAB architecture is based on this concept.

The Cloud RAN (C-RAN) concept was firstly introduced in the last two 3GPP releases of the standards for the 4th generation of mobile networks (Rel. 13, 14). Base stations were split into two units, the Remote Radio Head (RRH) and the Base Band Unit (BBU), while the RRH contained only the radio functionality and the BBU contained all the baseband processing. According to the C-RAN concept, the RRHs were placed on the antenna sites while the BBUs belonging to base stations of some specific area could be centralized in a remote site location. The connection between the two units of a base station was a PtP wireless connection, while this link was known as the fronthaul. This concept allowed the sharing of computing resources between BBUs and introduced important gains that have solidified C-RAN as one of the enabling technologies even for 5G. The centralized architecture reduces the operational and capital costs for the network operator, allows the flexible allocation of processing power to base stations that need it more and facilitates coordination between base stations for better interference mitigation and handovers [5].

The traffic volume supported by 5G networks increased dramatically compared to the 4th generation. Hence, some additional features had to be considered in C-RAN technology in order to become applicable in the next generation. The main problem was that the capacity demand on the fronthaul link became very high. This is why functional split came into the frame, since it defines how many base station functions will be hosted on the antenna site and how many of them will be hosted on the centralized pool. Of course, there is a trade-off introduced here. When functions are centralized, the network exploits all the benefits offered by C-RAN. On the contrary, when more functions are deployed near to the user, fronthaul capacity demand is decreased as well as the experienced latency. Eight different functional splits were initially proposed for examination by 3GPP in release 14 [6].These functional split options are depicted in Figure 2-3.

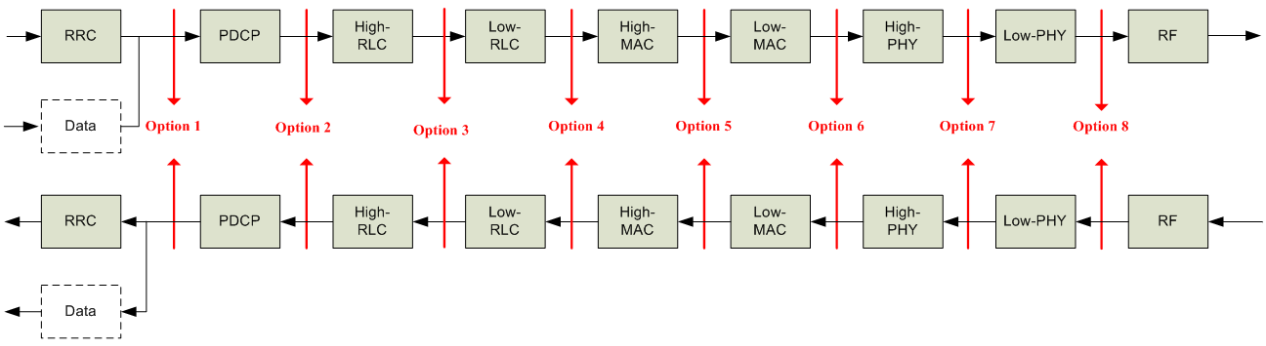

Figure 2-3:Functional split options between the centralized and distributed unit [6]

The second option of Figure 2-3 was later standardized in Release 15. According to the proposed New Generation - Radio Access Network (NG-RAN) architecture, illustrated in Figure 2-4, a base station (gNB) can be split into a Centralized Unit (CU) and one or more Distributed Units (DUs), where each DU hosts the Packet Data Convergence Protocol (PDCP) and Radio Resource Control (RRC) functions, while the rest of the functions are hosted in the CU [7].

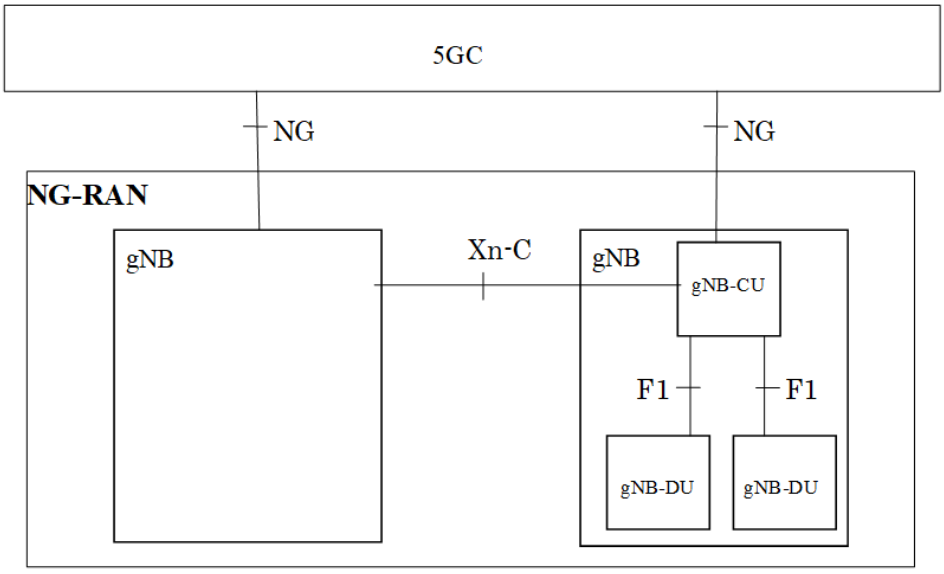

Figure 2-4:Overall NG-RAN architecture [7]

Moreover, an architecture with separated control and user plane functions is available (Figure 2-5), where a gNB might consist of a CU-Control Plane (CU-CP) and multiple CU-User Planes (CU-UPs) and DUs. The communication between CU and DU is realized through the F1 interface (F1-U for the user plane and F1-C for the control plane). IAB architecture is actually based on this standardized functional split scheme.

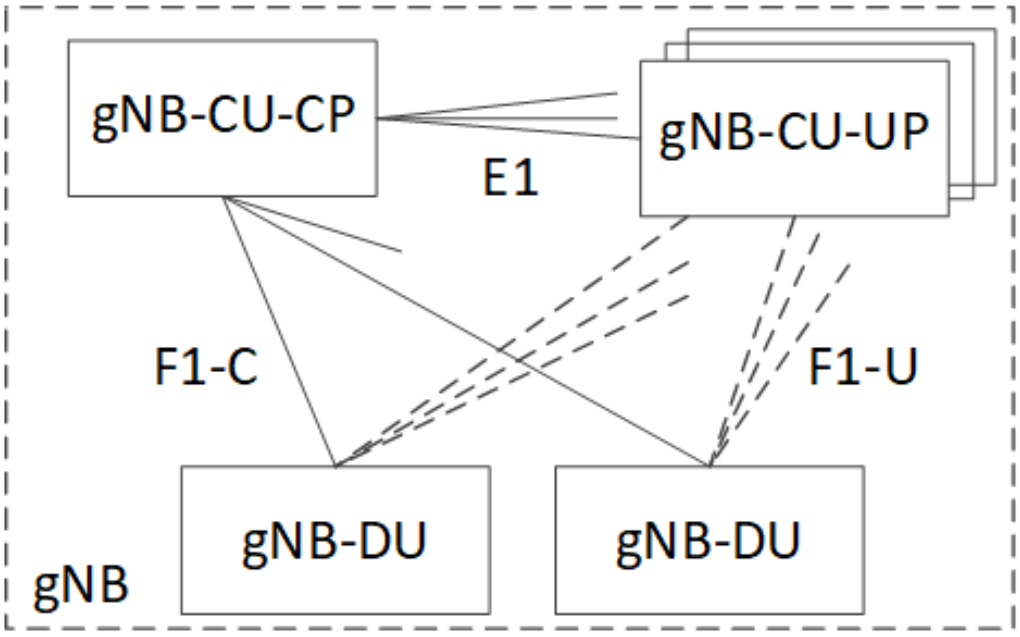

Figure 2-5:Overall architecture for separation of gNB-CU-CP and gNB-CU-UP [7]

2.2.2 IAB architecture

An IAB-node has to serve two different roles. On the one hand, it acts like a DU of a base station providing network access to UEs and other IAB nodes (this function of the node is called IAB-DU). On the other hand, its mobile terminal (IAB-MT) function supports a subset of the UE functionalities to enable connectivity with the IAB-DU of other nodes supporting IAB operation. An IAB-donor is a base station (gNB) connected with the Core network through a typical backhaul technology, and provides backhauling to UEs and IAB-nodes which are connected directly to it [8]. The overall architecture of IAB is given inFigure2-6, while the protocol stacks for the support of the F1-U and F1-C protocols are depicted in Figure 2-7 and Figure 2-8 respectively.

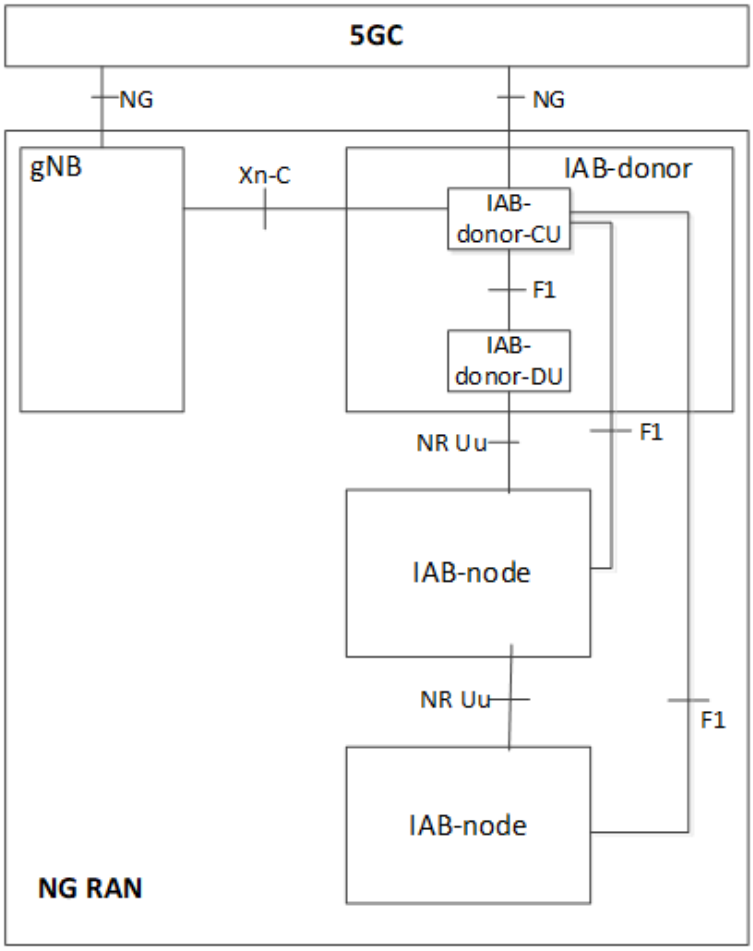

Figure 2-6:Overall IAB architecture [7]

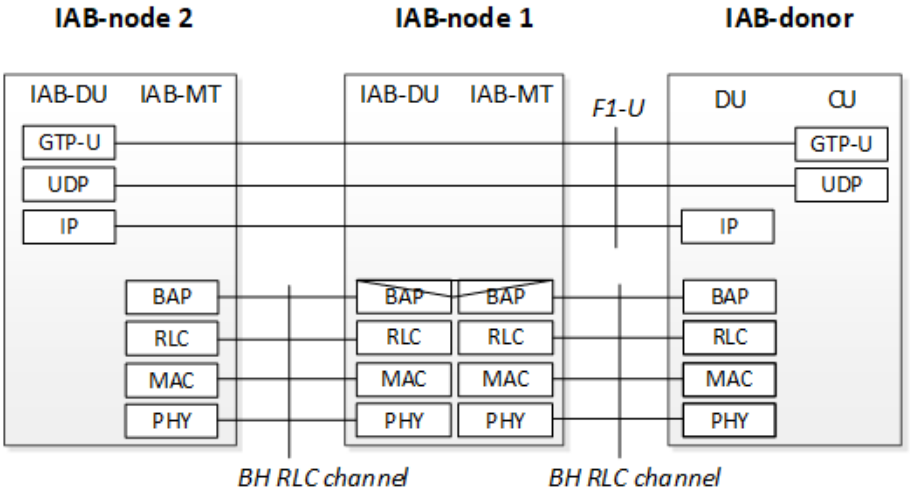

Figure 2-7:IAB protocol stack F1-U [8]

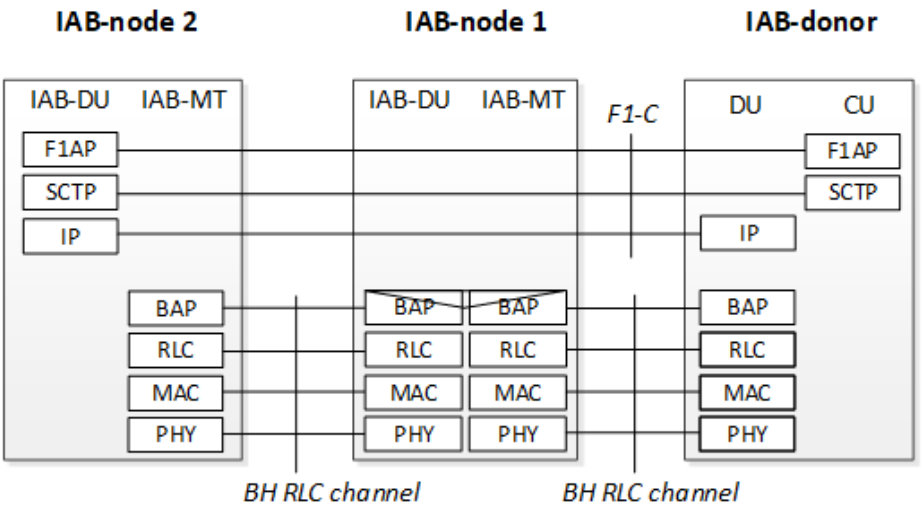

Figure 2-8:IAB protocol stack F1-C [8]

The IAB topology is a Directed Acyclic Graph (DAG), which comprises child and parent nodes. An IAB node is characterized as a child when its IAB-MT is connected to the IAB-DU of another IAB node or donor (called the parent). An illustration of this parent-child concept is given in Figure2-9.

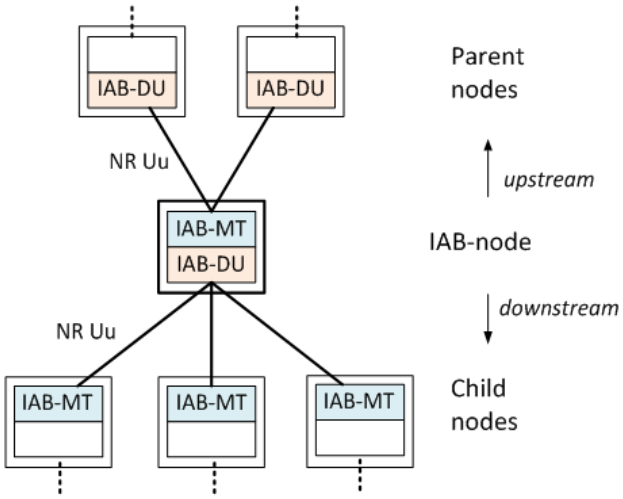

Figure 2-9:Parent and child IAB nodes [8]

2.2.3 Routing in IAB

A new protocol, called the Backhaul Adaptation Protocol (BAP), lies above the Radio Link Control (RLC) sublayer in the protocol stack and handles the forwarding of packets between IAB-nodes. At each IAB-node the BAP sublayer contains two collocated BAP entities, one at the IAB-DU function and another one at the IAB-MT function, while at the IAB-donor it contains only one (see Figure2-7). Also, each IAB entity has a receiving and a transmitting part to cater both for the uplink (UL) and the downlink (DL) [9].

The IAB-donor is responsible for assigning a BAP address to each one of the IAB-nodes and provides them with one routing table for the UL and one for the DL. In the UL, the IAB-node providing access to the UE inserts a BAP header with the BAP address of the IAB-donor-DU and an optional path ID in case there are multiple routing paths. Then each of the intermediate IAB-nodes makes a forwarding decision for the next hop according to the BAP header and the UP routing tables. The procedure is similar for the downlink, except that the IAB-donor now inserts the BAP header with the destination IAB node and the path ID, while each intermediate node looks up to the DL routing table. When the packet finally reaches the destination node, it is delivered to the upper layers of the protocol stack, after the BAP header has been removed first [4], [10].

After taking a routing decision, the BAP protocol has also to take the decision of mapping the ingress backhaul (BH) RLC channel to an egress BH RLC channel. There is the possibility of mapping only one ingress BH RLC channel to an egress BH RLC channel or multiplex more ingress channels into an egress one. This is done according to QoS criteria, while the procedure is again centrally configured, similarly to the routing [4], [10]. An example of how the BAP sublayer functions is given in Figure2-10.

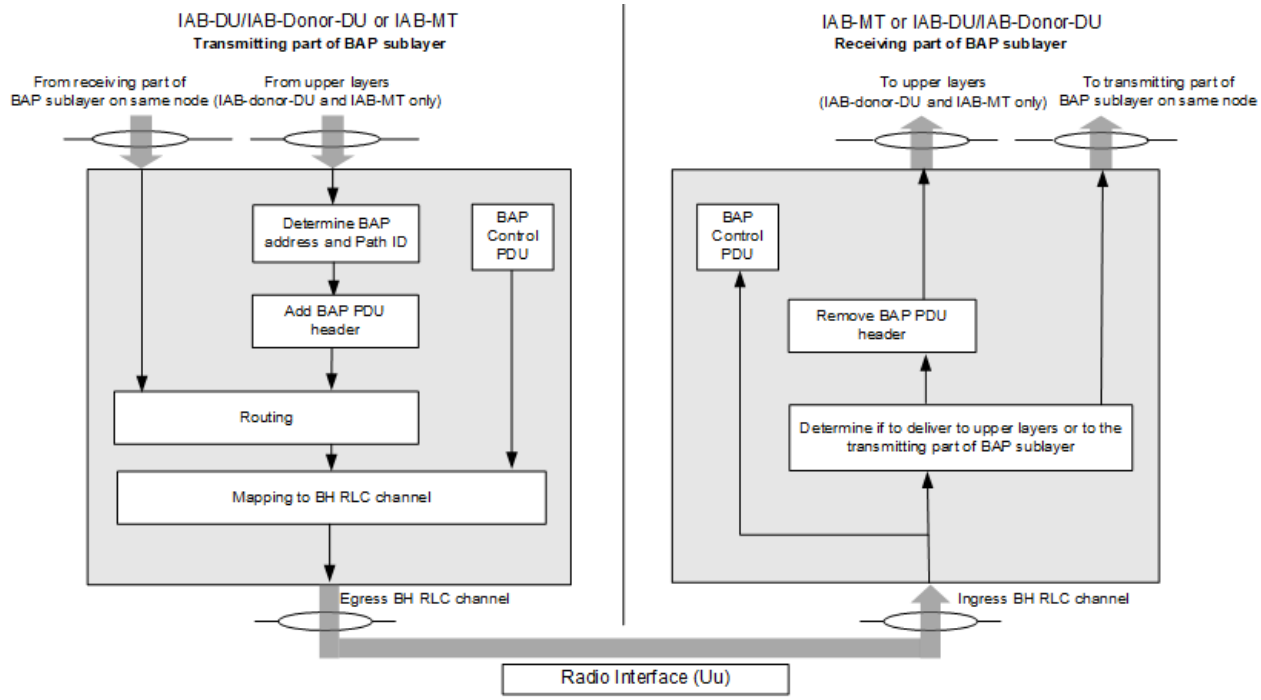

Figure 2-10:Example function of the BAP sublayer [9]

3 State of the art on end-to-end network slicing

3.1 Standard specific E2E slicing approaches

There are multiple standards organizations that define the standard specifications around network slicing. Several standardization body such as 3GPP, ETSI, ITU-T are working to establish key requirements for end-to-end network automation and orchestration. The 3GPP is the focal point for 5G Slicing Standards by conducting standardization effort holistically on network slicing, as considers network slicing as one of the key features for beyond 5G. It also defines four types of 5G slices: eMBB, URLCC, MMTC and V2X.

In 3GPP, a significant effort from several working groups is spent in order to develop comprehensive network slice related specifications such as,

- SA1 (service requirements), describes the use cases and requirements related to slices.
- SA2 (architecture) defines the core network slice infrastructure and control plane processes.
- SA3 (security) defines the slice security features.
- SA5 (network management) defines the slice management architecture.

Based on the above, 3GPP SA2 which is responsible for the overall system architecture, defines the relevant architectures and procedures of core network element support slice through a set of Technical Specifications (TS). These could be summarized to TS 23.501 (Stage-2 System Architecture for the 5G System which includes Network Slicing), TS 23.502 (procedures for the 5G System) and TS 23.503 (Policy and Charging Control Framework for the 5G System).

Respectively, 3GPP SA5 defines a series of specifications related to network slice management through also a set of TS. More specifically, TS 28.530 specifies network slice concepts, use cases and requirements. TS 28.531 defines the provisioning of network slicing for 5G networks and services. TS 28.541 defines the network resource model for network slicing and the SLA requirements related for end-to-end slicing are described by Service Profile, and the deployment resource requirements for slicing subnets are described by Slice Profile, covering parameters such as bandwidth, delay, and maximum UE number.

3GPP RAN defines the RAN side network slicing principles, slice selection, UE context processing, mobility and other signaling procedures in TS 38.300.

Finally, as previously mentioned, 3GPP TR 28.801 has introduced the network slice instance lifecycle management. The TR 28.801 is responsible for the management functions CSMF, NSMF, NSSMF.

It is highlighted that the 3GPP approach requires no modification of the ETSI MANO framework as the CSMF, NSMF and NSSMF functions are deployed within OSS/BSS part of the MANO

framework. However, the OSS/BSS internal functionalities are not defined by ETSI MANO and its architecture is presented in Figure 3-1.

The ETSI OSM NFV Orchestrator is leveraged for Network orchestration at the Core Cloud (i.e., service scale-out, scale-in and VNF placement). Being Open-Source software allows OSM to implement an ETSI-aligned NFV architecture, provide practical and essential feedback to the ETSI ISG NFV and increase the likelihood of interoperability among NFV implementations. OSM is a project adopted by ETSI, in an initiative to develop an Open Source NFV MANO software stack aligned with ETSI NFV. It implements the two key components of the ETSI NFV architectural framework, i.e., the NFV Orchestrator and VNF Manager, jointly known as NFV MANO. ETSI OSM also implements Network slicing Manager (i.e., acts as the CN NSSF) supporting a fully programmable northbound API, which follows the ETSI SOL-005 specifications. More specifically, ETSI OSM is capable of providing Network Slices as a service, assuming the role of Slice Manager as per ETSI NFV EVE012. Network Slices, in the OSM context, operate as a particular kind of Network Service or, more generally, as a set of various Network Services that are treated as a single entity.

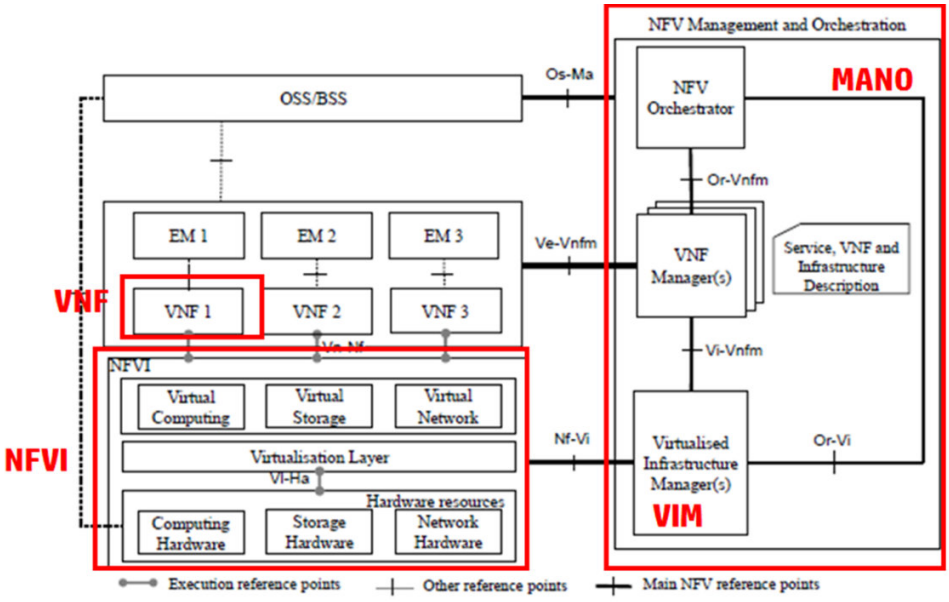

Figure 3-1:E2E Slicing Architecture (3GPP) [17]

3GPP defines the relation between Network Slices and Network Services (NS), where Network Services become the so-called NS subnets of the Network Slice, while the Network Slice with its constituent NS subnets can be deployed and operated as if they were a single entity. The Northbound Interface (NBI) of the OSM supports the allocation of resources to slices during the deployment stage and their adaptation to different usage conditions. The instantiation of a new Vertical application generally involves the creation of a new Network Slice Instance, or the sharing of an existing one. The slice creation request is sent to the OSM NBI, including the network slice requirements and the information indicating whether the requested NSI could be shared with other vertical applications.

3.2 AI/ML based E2E slicing approaches

AI based slice management framework is an emergent paradigm for slice management [11]. AI-enhanced solutions have been recently proposed in the literature, including supervised techniques, which includes ground truth data for training, unsupervised techniques, which work in absence of ground truth; and Reinforcement Learning (RL) approaches, where intelligent agents take actions in an environment, in order to maximize the so-called cumulative reward.

Reference [11] considers AI techniques for admission control. It considers the tradeoff between Key Performance Indicators (KPIs) fulfillment and resource sharing. Upon arrival of a new slice request, the system takes the action (i.e., accepting or rejecting the request) that maximizes long-term revenue; each NN is in charge of forecasting the revenue associated with one action.

Furthermore, it indicates that AI for resource orchestration is the key tradeoff is under provisioning and over dimensioning can be tackled by convolutional NN (CNN) architecture for time series prediction with a dedicated loss function as it allows exploiting inherent spatial correlations in the traffic generated at different geographical locations. [11] additionally, discusses about AI usage in Slice Scheduling at Radio Access part. The key challenge of network slicing is the design of a radio access network (RAN) virtualization (vRAN) mechanism that jointly provides isolation between network slices and adapts the allocation of pooled physical resources to the needs of each virtual RAN. A combination of unsupervised learning (deep auto-encoder) and deep reinforcement learning is a promising solution deep deterministic policy gradient (DDPG) algorithm, implemented by actor-critic NN structures, can deal with large and/or continuous action spaces, which are common in resource control problems.

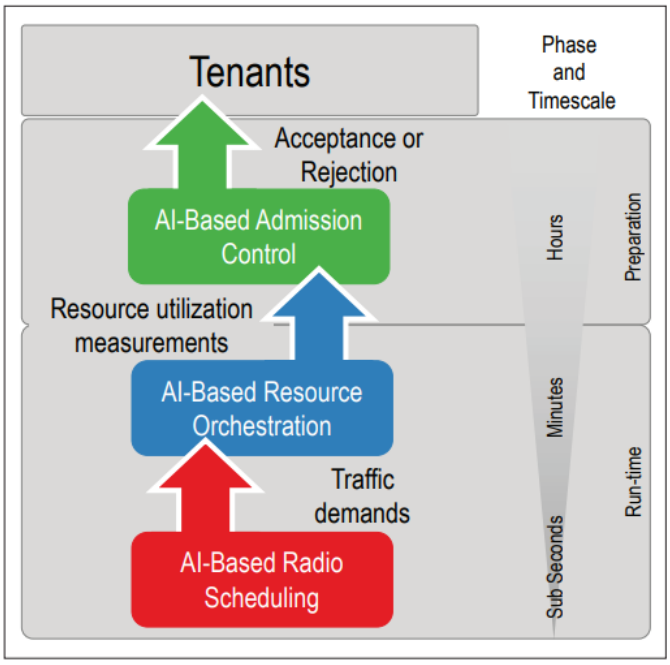

Figure 3-2:AI based slice management framework [11]

The data driven framework that effectively allocates capacity to individual slices by adopting an original multi-timescale forecasting model [11]. It uses deep Learning architectures and a traditional optimization algorithm and anticipates resource assignments that minimize the comprehensive management costs induced by resource overprovisioning, instantiation and reconfiguration, as well as by denied traffic demands

Isolation of resources across slices inherently increases network capacity requirements, and a dynamic, preemptive and efficient allocation of resources to slices. The use of AI can enable zero-touch networks, i.e., fully self-operating communication infrastructures where forecasting holds a fundamental role. To resolve issues such as,

- unnecessary resource overprovisioning,
- non-serviced demands,
- resource instantiation, and
- resource reconfiguration

The authors in above literature utilized concept of multi-timescale orchestration. The capacity forecasting tries to accommodate the demand and to limit overprovisioning, by reconfiguring resources proactively. This approach also increases the network slice admission rate and reduces subsequent re-configurations. A long-timescale orchestrator operates over extended intervals that span multiple re-orchestration opportunities; it allocates a dedicated capacity to each slice and also reserves an additional shared capacity accessible by any slice. Both capacities remain constant across the interval reducing resource instantiation [13].

Only the shared capacity is then reallocated at every re-orchestration opportunity by a short-timescale orchestrator, while the configuration of the dedicated capacity is preserved throughout the extended interval, thus reducing cost in terms of resource reconfiguration for both long- and short-timescale orchestrators.

The recent literature also discusses a deep neural network architecture inspired by advances in image processing and trained via a dedicated loss function returning a cost-aware capacity forecast, which can be directly used by operators to take short- and long-term reallocation decisions [20]. A deep learning architecture is utilized by authors in [20]. It exploits space- and time-independent correlations typical of mobile traffic and computes outputs at a data-center level which jointly solve the problem of capacity forecast in network slicing. It leverages a customized loss function that targets capacity forecast letting the operator tune the balance between overprovisioning and demand violations. Furthermore, [20] provides long-term forecasts over configurable prediction horizons, operating on a per-service basis in accordance with network slicing requirements. The work in paper aims at forecasting the (constant) capacity that should be allocated over a long-term horizon, so as to minimize the monetary cost incurred by the operator. At the core of [20], there is α-Operator Monetary Cost (alternative loss function), a new and

customized loss function that drives the deep neural network training so as to minimize the monetary cost contributed by two main deployment fees, i.e., overprovisioning and SLA violation.

The focus of [12] is on ML in wireless communications. 6G wireless communication networks will be the backbone of the digital transformation of societies by providing tackling ubiquitous, reliable, and near-instant wireless connectivity for humans and machines with advanced ML models, large datasets, and high computational power, Zero-touch optimization. Recent advances in ML research has led enable a wide range of novel technologies such as self-driving vehicles and voice assistants. Such innovation is possible as a result of the availability of advanced ML models, large datasets, and high computational power. On the other hand, the ever-increasing demand for connectivity will require a lot of innovation in 6G wireless networks, and ML tools will play a major role in solving problems in the wireless domain. In [12], authors provide an overview of the vision of how ML will impact the wireless communication systems. It discusses ML methods that have the highest potential to be used in wireless networks as well as the problems that can be solved by using ML in various layers of the network such as the physical layer, medium access layer, and application layer. This paper analyzes several AI-enhanced techniques for zero-touch wireless networks optimization, based on different approaches, such as for instance: Federated Learning, Deep Reinforcement Learning, Feed-forward DNNs, Kernel Hilbert Spaces and ML applied to PHY layer.

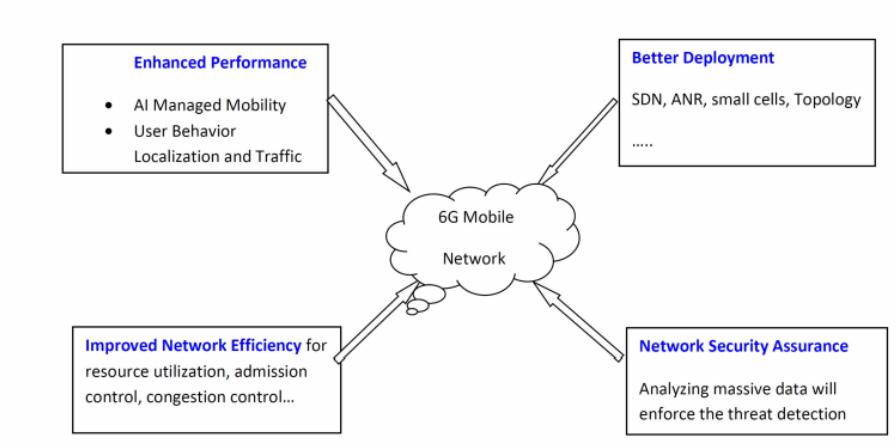

Figure 3-3: ML enhancing 6G network performance management [12]

3.3 Algorithms for efficient VNF placement and routing

One of the key problems that need to be resolved in order to fully leverage from the network slicing concept, is the optimal allocation of resources to slices in a virtualized environment. Efficient resource sharing is essential for hosting services with diverse Quality of Service (QoS) requirements over the same physical network infrastructure. Moreover, it can reduce the capital and operational costs for the network operators. Since network slicing is considered to be one of the key enablers of 5G and beyond networks, this problem has attracted the attention of the

research community and various optimization frameworks, models, and algorithms have been already proposed [21].

3.3.1 The network slicing problem description

A slice comprises a number of Virtual Network Functions (VNFs) chained together with a specified order. Within the concept of Network Slicing as a Service, a slice may even correspond to an end-to-end service. Such a service is represented by a graph called the VNF Forwarding Graph (VNF-FG) [22], and is usually also encountered as a Service Function Chain (SFC) [23]. Each VNF of a slice is associated with some computing, memory and networking requirements, and it can be instantiated on a server by reserving the respective resources. This server can be located either on the Core or on the Edge Cloud depending on the type of the VNF and the specific performance requirements of the slice. Moreover, adequate capacity should be reserved on the physical links that interconnect adjacent VNFs in a chain. Hence, the problem of resource allocation to slices in a virtualized environment translates into instantiating VNFs on servers and routing traffic between adjacent VNFs, making sure that some QoS constraints are respected for all slices.

A general optimization framework for the network slicing problem is given in [21]. In its simplest form, this problem can be formulated as an extended version of the Virtual Network Embedding (VNE) problem, which is already very well studied. Hence, the physical network is represented by a graph G=(V, E), where V is the set of physical nodes (SDN routers, servers or data centers) and E is the set of physical links connecting them. Each node and link has some specified capacity and an associated cost. In the simplest case scenario, the cost of using a node or a link is linear to the capacity reserved on it. On the other hand, a slice is a virtual network represented by H=(N, L), where N is the set of virtual nodes (VNFs) and L is the set of virtual links interconnecting them. Each virtual node and link have some specified capacity requirement and must be mapped to exactly one physical node and loop-free physical path respectively. A physical path consists of a number of physical links. The mapping of all nodes and links of a virtual network (slice) to the physical network is called a virtual network (slice) embedding and the objective is to find the embedding with the lowest cost. According to the problem constraints, the total capacity requirements of the virtual nodes/links embedded on a physical node/path cannot exceed its respective capacity. Moreover, each virtual node can be deployed only on a subset of V (only a subset of nodes can host VNFs, some VNFs can be instantiated only on a specific network domain). The VNE problem is formulated as an Integer Linear Program (ILP) which can be solved optimally but it is NP hard.

The basic formulation, outlined above, can be extended by utilizing different objectives, resources, constraints and network topologies to formulate any slicing problem. Various extended problems are encountered in the related work under different names, like the VNF-Forwarding Graph Embedding (VNF-FGE) problem [24] or the SFC embedding problem. Moreover, there are papers

where only the placement or only the routing problem is addressed individually. Another distinction between related work papers is that some of them address the offline problem, while others tackle the online one. In both of these approaches a number of slices must be embedded onto the network, however in the online counterpart some slices that have been already deployed on the network should be taken into account.

3.3.2 Models, objectives and constraints

The network slicing problem has been already examined in many research works. Consequently, a great variety of different models, objectives, constraints and optimization frameworks have been proposed. It is useful to outline some of them in selected related work papers found in the literature.

The service chain embedding problem is addressed in [25], by targeting to minimize the total service deployment costs for the network operator while at the same time providing QoS guarantees. The deployment costs refer to both operational and capital expenses and include the server cost, the cost for the network function licenses and the link bandwidth lease cost. The network is modeled by a graph comprising network nodes and links. Each service chain is a VNF-FG and is also associated with some specified data rate, maximum E2E latency and minimum availability. Moreover, VNFs require some specific amount of resources, which can be either processing, memory or storage and can handle only a limited amount of traffic, while incurring some processing delay. The constraints considered in this work, apart from the node and link capacities, include the minimum availability of service and the maximum end to end latency (both propagation and processing delay). The availability of service can be degraded due to a number of reasons like link failure, VNF software failure or server failure. Moreover, topological constraints are considered concerning the placement of VNFs. The authors formulate this problem as an ILP which is NP hard. Due to the fact that the problem is intractable for practical sized networks they also propose a greedy heuristic algorithm, which provides near-optimal solutions in reasonable execution time.

In [26] the authors formulate the Service Chain Embedding with Maximum Flow problem. They utilize a simple system model comprising only computational nodes interconnected by bi-directional links. They also consider a VNF-specific linear relation between the CPU or memory demand of a VNF and the traffic demand of the respective service. Each service chain is associated with a source-destination pair of nodes, some maximum amount of flow, and some processing demand per unit traffic demand. Moreover, each service chain is further divided into a number of flows, which can be routed through different links and nodes across the network, while each node that lies inside their path processes a fraction of the total demand. The objective is to maximize the total flow in the network while respecting the problem constraints. These concern the link capacity, the node capacity and the maximum flow associated with each service. The authors formulate the

problem as a Mixed Integer Linear Program (MILP). Since the original problem is NP-hard, an approximation algorithm is proposed, which is followed by a heuristic algorithm for solution improvement.

In [27], a new optimization framework that takes into account the relationship between efficient resource allocation and network resource pricing is proposed. The latter affects both the profit of the network operator and the profit of the service provider, since it has an impact on the demand for resources. The system is modeled by a weighted directed graph and comprises access nodes, forwarding nodes and data centers connected by links. Links and nodes have some specified bandwidth and processing capacity respectively. A slice is modeled as an ensemble of flows and each flow is associated with a source-destination node pair and a data rate demand. Moreover, a flow can be split even further and be steered towards multiple paths. A linear relationship between traffic demand and processing demand is also considered in this work. The objective is to maximize the net social welfare of the network, which is defined as the ratio of the total user utility to the cost of resources for the slice provider. A distributed low complexity algorithm is proposed for resource allocation that can achieve near-optimal net social welfare with profit guarantees.

The online routing problem is examined in [28], taking the VNF placement as an input. The objective is to minimize the maximum network congestion. The physical network is modeled by an undirected graph, while nodes and links are characterized by some capacity. A service chain is characterized by the source/destination nodes and the type and order of the associated VNFs. The traffic on the link between consecutive VNFs depends both on the traffic demand of the service and the type of VNFs, since they are considered to have traffic changing effects. The problem is formulated as an ILP and an efficient online algorithm inspired by the virtual circuit routing problem is proposed.

The Prioritized Service Function Chain Deployment problem is formulated in [29] as an MILP. Two types of chains are examined regarding their priority, the emergency and the best effort chains. On the one hand, the bandwidth requirements of the flows belonging to emergency chains must be satisfied. On the other hand, the rest of the flows share the remaining network resources provided that the emergency demands are guaranteed. The objective is to maximize the network provider's profit, which is a function of the income from the deployed chains taking also into account the energy consumption of the system. The income from a service chain is linearly related to the input traffic on this chain, while the energy costs of the servers are calculated based on a well-known power consumption model. Moreover, there is an additional term accounting for energy costs, since migrations of VNFs have also an impact on the energy consumption (it is higher during migration). Both an exact and a heuristic algorithm are proposed for the solution of this problem.

A different optimization framework, compared to the classic approach of placing each VNF to a node and interconnecting them, is proposed in [30].An architecture based on the cloud-native paradigm is considered here. According to this paradigm, each VNF can be further divided into

smaller components deployed across the cloud. These components, along with their interconnections, comprise a micro-slice, and then a slice consists of micro-slices instead of VNFs. This increased granularity allows much more flexibility in allocating resources. According to that model, a slice request can be represented by only one VNF with a total traffic volume, a corresponding processing requirement that is proportional to the traffic volume, and a pair of source-destination nodes. So, instead of trying to individually place and interconnect each VNF, the goal now is to distribute the transport and processing load over the network. The authors formulate a continuous optimization problem where the decision variables are the routing allocation variables (the amount of traffic of a slice routed through one of the paths and processed at one of the nodes, for all different slices, paths and nodes). The objective is to maximize the sum of an a-fair utility function over all slices. The solutions can be fair with respect to traffic, computing, or a mix of the two according to a slice-specific balancing parameter. The objective is to maximize the sum of an α-fair utility function over all slices, where the utility function represents the perceived worth of the allocated resources (traffic or processing) to a slice. Different values of α correspond to a different shape of the utility function. According to the model used in this paper, the solutions can be fair with respect to traffic, computing, or a mix of the two based on a slice-specific balancing parameter. Finally, the problem constraints concern the capacity of links and nodes.

From the related work we see that the physical network is usually modeled by a graph (undirected/directed/weighted). Most of the objectives are related in some way either with the operational or capital cost for the network operator or more generally with its profit. Also, in many cases the objective is linked with the performance of the slices/services. On the other hand, the constraints are usually related directly or indirectly with the QoS. Some of the QoS parameters used are latency and availability. Finally, traffic steering is usually utilized to alleviate the network from congestion.

3.3.3 Algorithms

The VNE problem, as well as any of its variants, is NP-hard. This holds even if we consider only the placement or the routing problem in isolation [31]. Hence, non-polynomial time is required for obtaining an optimal solution to any network slicing problem. Since this is possible only for very small-scale scenarios, in the majority of papers the authors formulate the problem as an ILP or MILP and then propose a heuristic algorithm that provides sub-optimal solutions in a shorter time scale.

Centralized algorithms have been mostly utilized so far to provide solutions in the slicing problem. However, there are some drawbacks in the centralized approach:

- Scalability issues
- Any failure in the centralized optimizer will lead to a complete failure in allocating resources

- Privacy issues for multi-domain resource allocation

This is why researchers started investigating the use of distributed algorithms for this task. In [30], the Alternating Direction Method of Multipliers (ADMM) algorithm is employed to decompose the slice resource allocation problem into three sub-problems: the slice owner problem, the cloud provider problem and the network provider problem. Each of these sub-problems has either a closed-form solution or can be solved in polynomial time. Hence, this distributed algorithm is able to provide an optimal solution to the global problem in practical execution time. A distributed scheme is also proposed in [32] for E2E resource allocation. According to this method, an auction is formed between the slices and the Data Centers (DCs). This means that each slice makes a bid for the resources it is interested for and based on the bids the DCs determine a price for each of the resources (and consequently determine the amount of resources allocated to each slice based on its bid). Then, each slice updates its offer by solving an optimization problem with an objective to maximize its utility and minimize the money paid. After a number of iterations the algorithm converges as the system reaches the Nash equilibrium. The solution provided by the distributed scheme is the same (optimal) with the one provided by the centralized scheme.

The network slicing paradigm has introduced some trade-offs, since high customization of services leads to reduction in resource sharing efficiency and increase in operational costs. This is due to the fact that slice traffic demands change dynamically over time, and hence an allocation of resources that is optimal in the present might be highly inefficient if some change in traffic occurs [33]. The dynamic and intelligent allocation of resources to slices is the only way to mitigate these costs. Machine Learning (ML) algorithms have been employed to solve or assist in solving the resource allocation problem recently [20], [34], [35]. The ability of ML algorithms to track traffic patterns and make a forecast for the slice traffic demands makes them a suitable candidate solution for this task.

3.3.4 Open research problems

The network slicing problem can be formulated as an extended VNE problem. Within this framework various models, objectives and constraints have been proposed in the literature so far for VNF placement and routing. The objectives are usually trying to capture the profit of the network operator, the QoS of the slice (in the benefit of the slice customer), or balance between the two. On the other hand, constraints are related with the maximum capacity of the network nodes and links, as well as QoS parameters (latency, availability, etc.). Despite the existing literature on this topic, there is still room to enhance the model and objectives used focusing on capturing the E2E slice performance in networks with heterogeneous resources (radio, CPU, memory, etc.).

The network slicing problem is NP-hard and therefore it is not possible to solve it optimally for practical scenarios of large networks within reasonable execution time. This is why heuristic

algorithms are usually proposed to provide sub-optimal solutions in shorter time scales. Still, the execution time of an algorithm is very important, since it has to be less than the reconfiguration period of the system. Moreover, the dynamic environment in network slicing dictates dynamic resource allocation facilitated by traffic prediction mechanisms. ML algorithms (centralized or distributed) have been recently employed for this task with success, but there are still a lot of open research challenges in this area. Reinforcement Learning (RL) algorithms can be considered as an appropriate solution, since they derive resource allocation policies which are near-optimal in the long run and have learning capabilities suitable for dynamic environments. However, any such algorithm should be also able to cope with the very large size of the state and action spaces in the network slicing problem. Towards this direction, RL approaches that incorporate some function approximation method (like a Deep Neural Network),in order to reduce the training time, could provide practical solutions.

3.4 E2E network and service management and orchestration

5G networks and beyond may consist of various resources and functions supplied by different vendors and deployed across multiple network segments, each typically conveying numerous technologies. Furthermore, to deliver end-to-end services to vertical customers, it may be expected that the underlay substrate span beyond one single service provider domain.

To manage this multi-domain, multi-vendor, and also supports several types of use cases such as URLLC, eMBB, mMTC in 5G network environment, autonomous management, and orchestration of the end-to-end (network and services at run-time is desired. Autonomous service management aims to replace human involvement with functions that automatically execute the operation, administration, and management (OAM) processes, which reduces OPEX (Operating Expenses) and improves efficiency.

3.4.1 Aspects of E2E network slicing management

The following end-to-end network slicing aspects should be considered from a management and orchestration perspective:

- Exposure capability support to the vertical industry.
- Life cycle management of network slicing.
- Performance management of network slicing.
- Fault management of network slicing.
- Analysis and collecting information of traffic and bandwidth of E2E network.

Hence, there is a need to expose service capabilities from all domains to the seamless integration of new vertical use cases and E2E automation management. Therefore, E2E Network slicing, which is one of the key technology enablers of 5G, can help in this regard.

3.4.2 Standardization of network slice management and orchestration

Several standardization bodies, such as 3GPP, ETSI (European Telecommunication Standards Institute) and ITU-T, are working to establish the key requirements for end-to-end network automation and orchestration.

Here, some of the most popular standards mention in this section.

- ETSI established the Zero Touch Network and Service Management Industry Specification Group (ZSM ISG) in 2017. ETSI ZSM ISG specifies a complete end-to-end network and service management reference architecture along with Artificial intelligence (AI) based closed-loop management automation and intent-based networking (IBN) approaches without requiring human intervention. The main aims of it to resolve the 5G E2E Network Slicing management issue.
- TMF's Zero-touch Orchestration, Operations, and Management (ZOOM) project [36] define a new management architecture of virtualized networks and services based on smooth interaction between physical and virtual components assemble into personalized services dynamically and enable same principles as ZSM, such as dynamic and open APIs (Application Programming Interfaces), closed-loop end-to-end management, near real-time, and zero-touch.
- ETSI ENI (Experiential Network Intelligence) ISG (Industry Specification Group) [37]defines a Cognitive Network Management architecture using closed-loop AI mechanisms based on context-aware and metadata-driven policies to improve the operator experience and defines different use cases that cover infrastructure management, network operations, service orchestration and management, and assurance.

3.4.3 Motivation of choosing ZSM

We plan to follow the ETSI ZSM framework to manage and orchestrate end-to-end network and services among the standards mentioned above.

Basically, The E2E network slicing solution procedure includes lifecycle management of network slice, provisioning, performance management and topology service, etc. ZSM framework supposes to be the best for an end-to-end network slicing and services management and orchestration. The reasons for choosing ZSM are exposed below:

- Manages different technological domains such as Core, RAN and Transport domains, and include managing several types of resources such as VNFs, SDNs, virtual and physical resources, etc.
- Focuses on the automation of E2E life cycle management of all the diverse types of network resources and services, including installation, commissioning,

configuration, day-2operations, software upgrades, and decommissioning. Figure 3.4 shows the way of coordination of AN, TN&CN management systems within ZSM framework.

- Includes the E2E management solutions for network slicing such as network slice cloning, isolation of network slices to ensure a sufficient level of independence between the net-work slice instances with tolerable interference, cross-domain network slicing management capability, etc.
- Supports vertical SLA (Service Level Agreement) management.

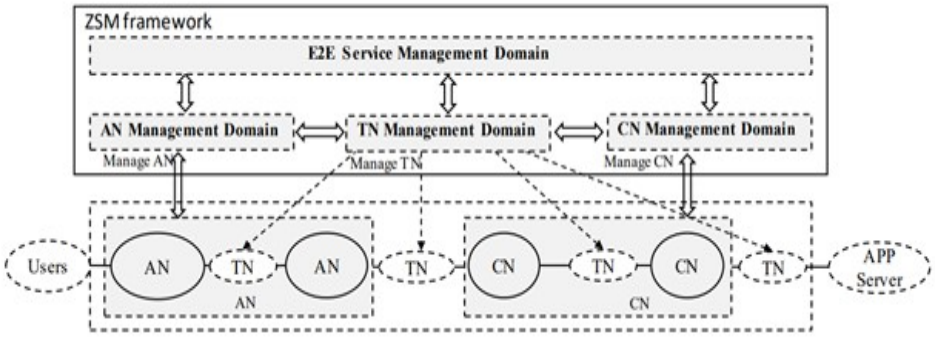

Figure 3-4:Coordination of AN, TN and CN management systems within ZSM framework [15]

3.4.4 Description of ZSM architecture

The ZSM framework designs the reference architecture to fully support network and service management in multi-domain, multi-technology environments. To meet this goal, the design of the ZSM architecture is modular, extensible, scalable, and resilient to failures. Modularity is a keystone for achieving the architecture extensibility, scalability, and resiliency, which collabo-rates with intent-based interfaces, closed-loop operation, and AI/ML techniques to empower the full-automation of the management operations. The ZSM framework designs the reference architecture to fully support network and service management in multi-domain, multi-technology environments. To meet this goal, the design of the ZSM architecture is modular, extensible, scalable, and resilient to failures.

As illustrated in Figure 3-5, the framework architecture comprises a set of architectural building blocks, namely, management domains (MDs), including E2E service MD, management services, integration fabric, and common data services.

Each MD is responsible for intelligent automation of orchestration, control, and assurance of resources and services within its scope. Resources can be physical resources (PNFs), virtual resources (VNFs), and cloud resources (X−as−a−service). It may contain domain data ser-vices that allow data sharing between functional components inside the MD. The E2E service MD is a particular MD that manages end-to-end, customer-facing services that span multiple domains provided by different administrative entities and coordinates between domains using orchestration.

Such kind of E2E service MD get away from monolithic systems, reducing the overall system's complexity and enabling independent evolution of domains and end-to-end management operations [38].

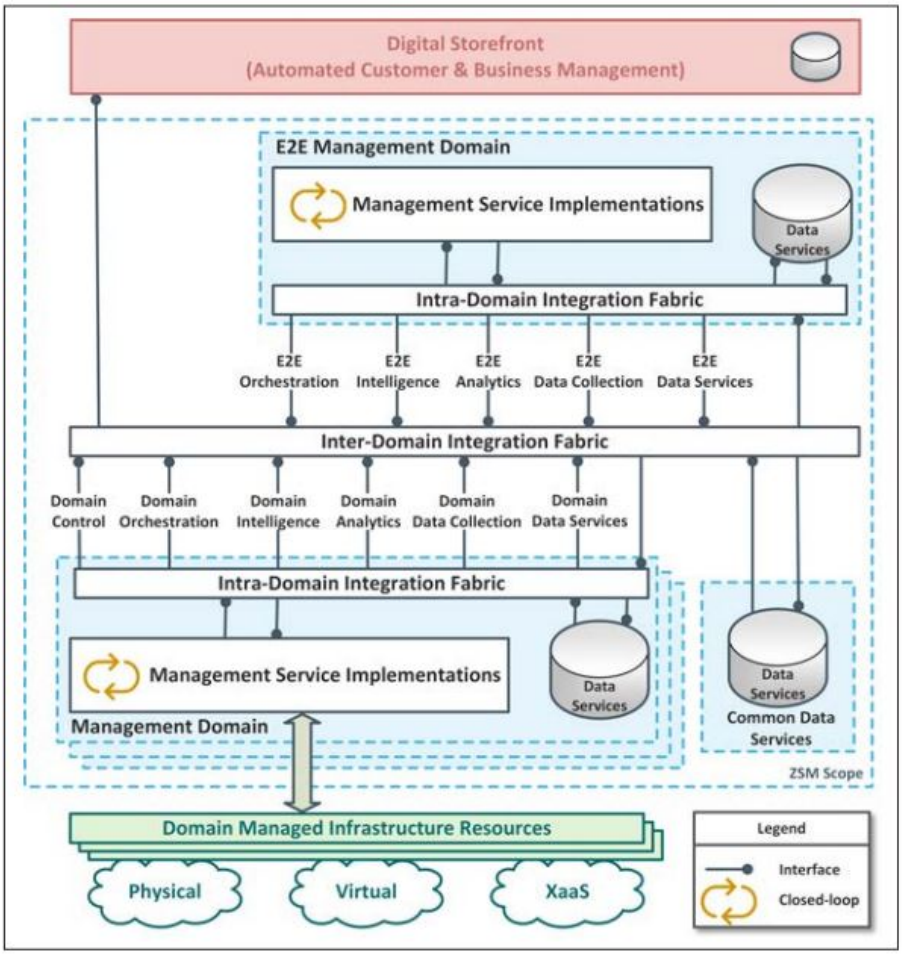

Figure 3-5:Reference architecture of ZSM [17]

The common data services allow separate data storage and data processing, facilitating access to data and cross-domain data exposure.MD domain and E2E service intelligence services use data stored in Common Data Services to drive domain-level and cross-domain AI-based closed-loop automation.

Finally, Each MD, including the E2E service MD, comprises several management functions grouped into logical groups (e.g., domain collection services, domain analytics services, domain intelligence services, domain orchestration services, and domain control services) and provides a set of management services via service interfaces. Some services offer and consume locally inside the domain using the intra-domain integration fabric [38].

3.4.5 Open research challenges

There are some open areas and related challenges for managing and orchestrate E2E network slicing in the ZSM perspective, which is needed to be solved shortly to get better efficiency and

performance. This section will mention some areas and associated challenges that may be our target to solve in the future.

- **VNF placement problem in ZSM**

Virtual Network Functions (VNFs)' autonomous placement is one of the critical aspects of ZSM in Fifth Generation (5G) networking. Most of the existing VNF placement work considers the only resource available for selecting a VNF host, neglecting other two aspects such as processing power and network performance, including the dynamic nature of network demand, which will affect the VNF placement efficiency. Moreover, another aspect that hinders ZSM is today's well-established orchestrators still call for human interaction for VNF placement-related work [39].

Hence, there are some open areas for exploring VNF placement from a distinct perspective: such as current orchestration frameworks that need to be enhanced, which will be

- The autonomous and intelligent one
- Adaptability to learns in operation according to the dynamic nature of the network
- Able to predict service-level end-to-end performance predictions accordingly.

Therefore, explore an appropriate ML technique that will successfully work for autonomous and intelligent VNF placement according to ZSM.

- **VNF/NS profiling**

Another vital area that will help to achieve the ZSM goals is called VNF/NS profiling system. A profiling system is nothing but the act of acquiring deep knowledge and create a mathematical or computational model about a computer-centric system, in our case, about NSs, VNFs, and VNFs chain. Such a profiling system helps to monitor the overall performance of the e2e network. For developing ZSM goals, there are some open areas which are needed to be considered-

- To acquire in-depth information about NSs and their constituent Virtual Network Functions (VNFs) to react proactively to the increase or decrease in demands, the next generation of intelligent NFV MANO systems is required.
- Additionally, various works have studied the integration and adaptation of NF/VNF profiler’s solutions to the existing NFV-MANO standard, and architectures are still awaiting [40].

- **End-to-End Network Service Monitoring**

The next-generation network will handle several types of services, use cases, and associate requirements. So, constant monitoring of an E2E network service's performance is essential to assure the desired conditions.

Some research directions related to this area mentions below [41].

- Recently, ETSI ZSM has not yet well-considered the overall service monitoring process for managing the entire lifecycle of the E2E network and services.
- Another aspect, the translation from intent-based services requirements into KPIs (Key Performance Indicators) and configurations for automated monitoring, is not yet a well-consolidated approach inside ZSM.

- **ML Collaboration across Management Domains**

ZSM architecture comprises several management domains (MDs), including E2E service MD, management services, integration fabric, and shared data services. Each MD's responsibility is the automation of orchestration, control, and assurance of resources and services intelligently within its scope. The E2E service MD manages end-to-end, customer-facing services that span multiple domains provided by different administrative entities.

In that place, some future research directions are still open.

- Domain-specific knowledge and the aggregation of different analytics models across management domains are essential issues that are still open for future research.
- Using distributed federated learning approach across the MD faces some challenges such as expensive communications, system diversity, statistical heterogeneity, and privacy concerns [42].
- Designing a suitable management architecture that can fully leverage ML capabilities for cross-domain management while meeting the heterogeneity and scalability requirements becomes a fundamental problem for future research [43].

Above mentioned areas are inter-connected with each other. Such as proper modeling of NF/VNF profiler can help later for both service monitoring and VNF placement, which will increase the overall efficiency and performance of E2E network and services. Soon, we have planned to work on this related area using AI/ML-enabled ZSM concepts such as closed control automation and intent-based networking.

4 Conclusions

This document provided the state of the art on IAB and E2E slicing for 5G and beyond networks. The first section was an introduction to the topic, including the main architectural elements in E2E slicing and IAB according to the latest 3GPP specifications. The second section focused mainly on E2E slicing, providing state of the art information on recent standardization efforts as well as on research outputs, and identified some open research challenges. More specifically, 3.1 provided more details on the working groups and standardization efforts of 3GPP regarding E2E slicing, 3.2

outlined some current research efforts in using AI/ML methods to assist E2E slicing, 3.3 included a survey on the algorithms and optimization frameworks for VNF placement and routing, and finally 3.4 discussed standardization efforts on network slice management and orchestration, described the ETSI ZSM framework, and outlined some open research challenges.

5 References

[1] 3GPP, "Technical Specification Group Services and System Aspects; 5G; Management and orchestration Concepts, use cases and requirements," TS 28.530, 2018.

[2] 3GPP, "Provisioning of network slice for 5G network and services," TS 28.531.

[3] H. Ronkainen, J. Edstam, A. Ericsson and C. Ostberg, "Integrated access and backhaul - a new type of wireless backhaul in 5g," *Ericsson Technology Review,* 2020.

[4] C. Madapatha, B. Makki, C. Fang, O. Teyeb, E. Dahlman, M. Alouini and T. Svensson, "On Integrated Access and Backhaul Networks: Current Status and Potentials," *IEEE Open Journal of the Communications Society,* pp. 1374-1389, 2020.

[5] L. M. P. Larsen, A. Checko and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Communications Surveys & Tutorials,* pp. 146-172, 2019.

[6] 3GPP, "Study on new radio access technology: Radio access architecture and interfaces," TR 38.801, 2016.

[7] 3GPP, "Ng-ran; architecture description," TS 38.401, 2017.

[8] 3GPP, "Nr; nr and ng-ran overall description; stage-2," TS 38.300, 2017.

[9] 3GPP, "Nr; backhaul adaptation protocol (bap) specification," TS 38.340, 2019.

[10] O. Teyeb, A. Muhammad, G. Mildh, E. Dahlman, F. Barac and B. Makki, "Integrated access backhauled networks," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019.

[11] D. Bega, M. Gramaglia, M. Fiore, A. Banchs and X. Costa-Perez, "Network Slicing Meets Artificial Intelligence: An AI-Based Framework for Slice Management," *IEEE Communications Magazine,* pp. 32-38, 2020.

[12] A. Amad, W. Saad, N. Rajatheva et al., "6G White Paper on Machine Learning in Wireless Communication Networks," 28 April 2020. [Online]. Available: https://arxiv.org/pdf/2004.13875.pdf.

[13] D. Bega, M. Gramaglia, A. Garcia-Saavedra, M. Fiore, A. Banchs and X. Costa-Perez, "AZTEC: Anticipatory Capacity Allocation for Zero-Touch Network Slicing," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020.

[14] 3GPP, "Stage-2 System Architecture for the 5G System which includes Network Slicing," TS 23.501.

[15] 3GPP, "Procedures for 5G system," TS 23.502.

[16] 3GPP, "Policy and Charging Control Framework for the 5G System," TS 23.503.

[17] 3GPP, "Specifies network slice concepts, use cases and requirements," TS 28.530.

[18] 3GPP, "Network resource model for network slicing and SLA requirement related to E2E slicing," TS 28.541.

[19] 3GPP, "RAN ntework slivcoing principles, slice selection, UE context processinf, mobility and other procedures," TS 38.300.

[20] D. Bega, M. Gramaglia, M. Fiore, A. Banchs and X. Costa-Perez, "Deepcog: Cognitive network management in sliced 5g networks with deep learning," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019.

[21] S. Vassilaras, L. Gkatzikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbah and G. S. Paschos, "The Algorithmic Aspects of Network Slicing," *IEEE Communications Magazine,* pp. 112-119, 2017.

[22] ETSI, "Network functions virtualization (NFV); architectural framework v1.1.1," ETSI GS NFV 002, 2013.

[23] J. M. Halpern and C. Pignataro, "Service Function Chaining (SFC) Architecture," 2015.

[24] F. Schardong, I. Nunes and A. Schaeffer-Filho, "Nfv resource allocation: a systematic review and taxonomy of vnf forwarding graph embedding," *Computer Networks,* 2021.

[25] P. Vizarreta, M. Condoluci, C. M. Machuca, T. Mahmoodi and W. Kellerer, "Qos-driven function placement reducing expenditures in nfv deployments," in *IEEE International Conference on Communications (ICC)*, 2017.

[26] J. Kuo, S. Shen, H. Kang, D. Yang, M. Tsai and W. Chen, "Service chain embedding with maximum flow in software defined network and application to the next-generation cellular network architecture," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 2017.

[27] G. Wang, G. Feng, W. Tan, S. Qin, R. Wen and S. Sun, "Resource Allocation for Network Slices in 5G with Network Resource Pricing," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017.

[28] L. Gao and G. N. Rouskas, "On congestion minimization for service chain routing problems," in *ICC 2019 - 2019 IEEE International Conference on Communications*, 2019.

[29] B. Farkiani, B. Bakhshi, S. A. MirHassani, T. Wauters, B. Volckaert and F. De Turck, "Prioritized deployment of dynamic service function chains," *IEEE/ACM Transactions on,* pp. 1-15, 2021.

[30] M. Leconte, G. S. Paschos, P. Mertikopoulos and U. C. Kozat, "A resource allocation framework for network slicing," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018.

[31] M. Rost and S. Schmid, "On the hardness and inapproximability of virtual network embeddings," *IEEE/ACM Transactions on Networking,* vol. 28, no. 2, pp. 791-803, 2020.

[32] H. Halabian, "Distributed resource allocation optimization in 5g virtualized networks," *IEEE Journal on Selected Areas in Communications,* vol. 37, no. 3, pp. 627-642, 2019.

[33] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs and X. Costa-Perez, "How should i slice my network? a multi-service empirical evaluation of resource sharing efficiency," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking - MobiCom '18*, New York, 2018.

[34] C. Zhang, M. Fiore, C. Ziemlicki and P. Patras, "Microscope: Mobile service traffic decomposition for network slicing as a service," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking - MobiCom '20*, New York, 2020.

[35] P. T. A. Quang, A. Bradai, K. D. Singh and Y. Hadjadj-Aoul, "Multi-domain non-cooperative vnf-fg embedding: A deep reinforcement learning approach," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019.

[36] "Zoom-project," [Online] https://www.tmforum.org/collaboration/zoom-project/.

[37] [Online] https://www.itu.int/en/ITU-T/ focusgroups/ml5g/Pages/default. Aspx.

[38] ETSI GS ZSM 002, Aug. 2019.

[39] M. Bunyakitanon, X. Vasilakos, R. Nejabati and D. Simeonidou, "End-to-End Performance-Based Autonomous VNF Placement With Adopted Reinforcement Learning," *IEEE Transactions on Cognitive Communications and Networking,* vol. 3, no. 2, pp. 534--547, 2020.

[40] S. Moazzeni et al., "A Novel Autonomous Profiling Method for the Next Generation NFV Orchestrators," *IEEE Transactions on Network and Service Management,* vol. 18, pp. 642-655, 2021.

[41] N. Saraiva, D. Lachos, CE. Rothenberg, PH. Gomes, "End-to-End Network Service Monitoring for Zero-Touch Networks".

[42] T. Li, A. K. Sahu, A. Talwalkar and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine,* vol. 37, pp. 50-60, 2020.

[43] Q. Duan, "Intelligent and Autonomous Management in Cloud-Native Future Networks—A Survey on Related Standards from an Architectural Perspective," *Future Internet,* vol. 13, p. 42, 2021.

[44] "Zero-touch Network and Service Management," [Online]https://portal.etsi.org/TBSiteMap/ZSM/OperatorWhitePaper, Dec, 2017.