Received 14 March 2022; accepted 1 April 2022. Date of publication 6 April 2022; date of current version 18 April 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3165000

Mapping the VNFs and VLs of a RAN Slice Onto Intelligent PoPs in Beyond 5G Mobile Networks

MOHAMMAD ASIF HABIBI [1], FAQIR ZARRAR YOUSAF [1] (Member, IEEE), AND HANS D. SCHOTTEN [1,3] (Member, IEEE)

[1] Division of Wireless Communications and Radio Navigation, Department of Electrical and Computer Engineering, Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany

[2] 6G Networks Group, NEC Laboratories Europe GmbH, 69115 Heidelberg, Germany

[3] Intelligent Networking Research Group, German Research Center for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany

CORRESPONDING AUTHOR: M. A. HABIBI (e-mail: asif@eit.uni-kl.de)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Program through Hexa-X under Grant 101015956, and in part by the 5G-CARMEN Project under Grant 825012.

ABSTRACT The mapping of a virtual network service onto a physical network infrastructure is a challenging task due to the joint allocation of virtual resources across nodes and links, the diverse technical requirements of end-users, the coordination between multiple host domains, and others. This issue is exacerbated further by the extension of virtualization to the next-generation radio access network (NG-RAN) architecture and the provisioning of radio access network (RAN) slicing. To that end, this article focuses on the mapping problem of the virtual network functions (VNFs), as well as their internal and external virtual links (VLs), of a RAN slice subnet onto intelligent points of presence (I-PoPs) and transport networks in the NG-RAN architecture. In this context, in contrast to the majority of the state-of-the-art proposals, which frequently fail to achieve performance objectives and neglect resource allocation constraints, this article introduces automation and intelligence at an architectural level to map VNFs and VLs onto their corresponding physical nodes and links, with the goal of achieving superior efficiency in virtual resource utilization while granting the performance of a RAN slice subnet. Benefiting from a top-down approach, the key contributions of this article are: *(i)* to extend the architectural framework of network slicing towards the NG-RAN architecture and provide a comprehensive overview and critical analysis of the components and functionalities of a RAN slice subnet; *(ii)* to integrate the Experiential Network Intelligence (ENI) framework into a joint architecture of the network functions virtualization–management and orchestration (NFV–MANO), Third Generation Partnership Project-network slicing management system (3GPP-NSMS), and I-PoPs in order to render automation and intelligence to the management and orchestration aspects of a RAN slice subnet in the NG-RAN architecture; and *(iii)* to propose a learning-assisted architectural solution for mapping the VNFs, as well as their internal and external VLs, of a RAN slice subnet onto the underlying I-PoPs and transport networks.

INDEX TERMS 5G, automation, beyond 5G, intelligence, management and orchestration, mapping, network slicing, NG-RAN architecture, RAN slicing, resource allocation, virtual links, virtual network functions, virtual resources, virtualization.

I. INTRODUCTORY REMARKS

THE THIRD Generation Partnership Project (3GPP) has defined that an end-to-end (E2E) network slice (NS) is composed of two network slice subnets (NSSs): the core network (CN) NSS and the radio access network (RAN) NSS [1]. Both NSSs must be connected via a highly reliable and high-capacity transport network, which goes beyond the scope of the 3GPP. However, within the scope

of the Fifth Generation Public-Private Partnership (5GPPP), projects such as 5G-Crosshaul [2], 5G-NORMA [3], 5G-ESSENCE [4], and 5G-CHARISMA [5] addressed network slicing in transport network. The 5G-Crosshaul, in particular, proposed a transport network NSS in [6]. By chaining the network functions (NFs), pairing the network links, and harmonizing the allocation of network resources of the CN, RAN, and transport network NSSs; an E2E NS is realized [7]. Throughout its lifetime, the E2E NS must fulfill the requirements of the requested communication services of a communication service customer (also known as tenant) [8]. A tenant could be a mobile virtual network operator (MVNO) or the owner of a business in a service sector such as automotive, health care, agriculture, and many others.

In the 3GPP-defined next-generation radio access network (NG-RAN) architecture, each RAN NSS (hereinafter referred to as a RAN slice) could be composed of virtual network functions (VNFs) and physical network functions (PNFs) [9]. The NG-RAN architecture consists of a number of next-generation NodeBs (gNBs). Each gNB is composed of a single centralized unit (CU), a minimum of one distributed unit (DU), and a minimum of one radio unit (RU) [10]. These components (further details on them can be found later in this article) accommodate and distribute the VNFs and PNFs of a RAN slice in a flexible manner [11]. The PNFs – which are executed on top of dedicated hardware units and require physical resources such as transmission and receiving antennas, control units, digital signal processors, and others – are allocated, configured, and managed by the 3GPP-network slicing management system (3GPP-NSMS) entities [1]. The VNFs – which are deployed on top of general-purpose hardware units and require virtual resources such as virtual compute, networking, and storage – are configured and managed by the network function virtualization-management and orchestration (NFV-MANO) functional blocks (FBs) [12], defined by the European Telecommunications Standards Institute (ETSI). To manage and orchestrate the VNFs, PNFs, virtual resources, and physical resources of an E2E NS effectively, the ETSI and 3GPP jointly proposed a unified framework of the NFV-MANO FBs and 3GPP-NSMS entities in [12] and [13], respectively. This framework can be extended towards the NG-RAN in order to design an interoperable functioning architecture for managing and orchestrating the physical and virtual components of a RAN slice [9].

In order to employ the joint framework of the 3GPP-NSMS and NFV-MANO in the context of RAN slicing in the NG-RAN architecture, the VNFs and PNFs of a gNB must be defined in the first place [14]. To that end, we consider the CU and DU to be VNFs, and the RU to be a PNF in a RAN slice. Based on these assumptions, which are also in compliance with the philosophy of the Open-RAN (O-RAN) Alliance's RAN architecture [15], we refer to CU and DU as virtual centralized unit (vCU) and virtual distributed unit (vDU) in this article, respectively. Following that, the vCU, vDU, and RU must be dynamically mapped onto the aggregation data center, edge data center, and cellular network site, respectively, using an efficient mapping algorithm [16].

Each data center, regardless of its location on the edge or in the core of a cellular network, is also referred to as the network function virtualization infrastructure point of presence (NFVI-PoP) in ETSI terminology. The NFVI-PoPs are distributed geographically across the underlying cloud infrastructure, which is owned by a network infrastructure provider and could be managed by multiple service providers. Each NFVI-PoP is made up of physical nodes that can be virtualized to create a number of virtual nodes. To meet the low latency and high bandwidth requirements of the end-users, the NFVI-PoPs are typically interconnected via optical fiber, forming an E2E NFVI-PoP interconnection network. The physical paths in the underlying optical networks can also be virtualized into numerous virtual paths. The cellular network site is defined as a collection of nodes and networking equipment that are used to host the functionalities and links of the RU in a physical manner.

The primary goal of the mapping process is to deploy the VNFs and PNF of a RAN slice on suitable virtual and physical nodes that are capable of fulfilling their resource requirements in NFVI-PoPs and cellular network site, respectively. The vCU and vDU require virtual resources that are abstracted from underlying general-purpose hardware units in aggregation and edge data centers, respectively. The RU requires physical resources, which are available at a cellular network site. To construct a complete RAN slice, the virtual resources of vCU and vDU, as well as the physical resources of the RU, must be dynamically and orderly chained together, ensuring that network traffic is transported in both upstream and downstream directions [17]. Furthermore, such a service chain comprises internal and external virtual links (VLs) as well as internal and external physical links (PLs) that are used to connect the vCU, vDU, and RU in a RAN slice. The VLs and PLs must also be efficiently mapped onto their corresponding virtual and physical environments in the underlying transport networks, respectively.

The mapping of the VNFs and VLs of a RAN slice, which is the main focus of this article, is accomplished in several operational stages. In the beginning, the service chain of a RAN slice must be designed. To gain a better understanding of the order of its components, the internal functionalities of vCU and vDU, as well as the internal and external VLs of a gNB, must be specified and linked sequentially in each RAN slice. Thereafter, the number of virtual resources required by the vCU, vDU, and VLs should be quantified dynamically. Following that, the appropriate virtualized environments must be selected to accommodate the virtual resource requirements of a RAN slice. Afterward, the VNFs and VLs are mapped onto their corresponding virtualized environments in NFVI-PoPs and transport networks, respectively. Finally, the allocated virtual resources must be constantly optimized – using cutting-edge optimization models and resource prediction techniques – in order to

improve resource allocation, reduce energy consumption, lower capital expenditure (CAPEX) and operational expenditure (OPEX), and enhance the performance of the RAN slice in the NG-RAN architecture.

A. LITERATURE REVIEW

1) THE FOUNDATION FOR VIRTUALIZATION AND NETWORK SLICING IN THE NG-RAN ARCHITECTURE

To this end, a significant number of standardization efforts, industrial experiments, and theoretical research have been undertaken for the sake of exploring virtualization of NFs and network links, the allocation of virtual and physical resources, and the management and orchestration of the VNFs and PNFs from the core network down to the RAN architecture. Specifically, following the initiative by the Next Generation Mobile Networks (NGMN) Alliance for the foundation and definition of network slicing in fifth-generation (5G) mobile communication systems [18], several standards development organizations (SDOs) – such as the 3GPP, the ETSI, the Global System for Mobile Communications Association (GSMA), the O-RAN Alliance, and the Internet Engineering Task Force (IETF) – and their respective members from the industry have defined their visions and provided guidelines to concretize the concept of network slicing towards its realization in the NG-RAN architecture.

The scope of each SDO varies. For example, the NGMN Alliance has primarily concentrated on defining and developing key concepts for E2E network slicing that are also applicable to slicing the NG-RAN [18], the requirements of a RAN slice, and the architecture that allocates resources for a RAN slice [19]. To delve deeper into RAN slicing, the NGMN Alliance has recently studied the lower and higher layer split options of the functionalities of a RAN slice that might be hosted by the underlying infrastructure in a centralized or distributed manner [20]. These multiple options of functional split require connectivity over the NG-RAN architecture. To that aim, the NGMN Alliance has additionally explored a number of deployment possibilities for the transport network that connects the components of a RAN slice [21].

The 3GPP has been at the forefront of the standardization of RAN architecture since several decades. However, with the extension of virtualization to the NG-RAN architecture and the deployment of RAN slicing, a number of new aspects and features have emerged that fall outside its scope [14]. Thus, the 3GPP specifications have been focusing merely on *(i)* the management and orchestration of physical resources of a RAN slice [1], [13]; *(ii)* defining the requirements of communication services of various type of RAN slices [22]; and *(iii)* the distribution of the functionalities of a RAN slice over the NG-RAN architecture [23].

In contrast to the 3GPP, the ETSI-related industry specification groups (ISGs) – particularly the Network Function Virtualization (NFV) and Next Generation Protocols (NGP) – have been solely focused on the virtualization aspects of RAN slicing. These ISGs have specified, among other aspects, the management and orchestration of the VNFs and VLs, as well as the virtualization of underlying physical resources and the allocation of virtual resources for RAN slices. Furthermore, the ISG NFV proposed an architectural framework that can be integrated with an operations support system/business support system (OSS/BSS) [12], [24]. To effectively manage the underlying resources and infrastructure, the ISG NFV has specified several FBs that employ machine-processable description files to automate the operations of its proposed architecture [25].

Despite the fact that the ISG NFV promises to automate the operations of a RAN slice, many of the tasks are still performed manually by the network administrator and/or tenant. As a result, full dynamism, E2E automation, intelligence, and high scalability will be critical to the success of NFV in the next generation of mobile networks. To accomplish this, the ETSI has recently established the Experiential Networked Intelligence (ENI) ISG, which focuses on the implementation of artificial intelligence (AI) techniques, notably machine learning (ML) algorithms, in the NFV architecture and OSS/BSS for 5G and beyond communication systems [26]. The ENI System collects a large collection of performance-related metrics from the NFV-MANO, 3GPP-NSMS, and NFVI-PoPs autonomously in order to understand their configuration and operational status in real-time and then employs ML algorithms to enable intelligent service deployment, resource management, monitoring, maintenance, predictions, and other operations. The grand objective of the integration of intelligence and automation into the aforementioned systems is to improve efficiency in network operation, enhance the performance of the RAN slice, automate complex human-dependent decisions and processes, and many others.

The O-RAN Alliance has been tackling both the physical and virtual aspects of NG-RAN with the goal of transforming the traditional RAN (also referred to as vendor lock-in RAN) into an intelligent, virtualized, slicing-aware, and multivendor interoperable architecture that must operate based on open and disaggregated interfaces [27]. The key innovations introduced by the O-RAN Alliance are: (a) standardizing an open interface between the DU and RU, which is critical for lowering the total cost of RAN deployment and eliminating proprietary lock-in; (b) proposing the near real-time RAN Intelligent Controller (nrt-RIC) network optimization entity for controlling the components and resources of a RAN slice in NFVI-PoPs; and (c) introducing a non-real-time RAN Intelligent Controller (non-RT RIC) intent-based management entity for realizing automation and intelligence (notably ML algorithms) in all levels of the management and orchestration aspects of the open NG-RAN architecture [15], [28].

The GSMA has identified a large number of industrial use cases that require network slicing solutions, as well as the functional, operational, and performance requirements associated with them [29]. It has also proposed the

TABLE 1. A comparison of our proposed architectural solution to the most recent state-of-the-art works.

Reference	Comparative parameters					
	Distribution of vCU/vDU	Considered resources	Type of RAN	AI/ML-driven	No. of domains	Core objectives
[10]	Not a concern	Networking	NG-RAN	No	Multi-domains	Middlehaul latency
[11]	Dynamic	Physical resource blocks	NG-RAN	No	Not a concern	Functional split
[42]	Static	Compute	C-RAN	No	Multi-domains	VNFs deployment
[48]	Static	Not a concern	NG-RAN	No	Multi-domains	Description files
[49]	Dynamic	Not a concern	O-RAN	Yes	Multi-domains	Functional split
[50]	Static	Compute and storage	C-RAN	No	Multi-domains	RAN slice mapping
[51]	Dynamic	Compute	NG-RAN	No	Multi-domains	VNFs placement
[52]	Static	Compute, storage, and networking	C-RAN	No	Multi-domains	Scheduling VNFs in RAN
[53]	Static	Networking	NG-RAN	No	Multi-domains	Fronthaul delay
[54]	Not a concern	Compute	E2E	No	Not a concern	VNFs embedding
Our proposal	Dynamic	Compute, storage, and networking	NG-RAN	Yes	Multi-domains	VNFs and VLs mapping

Generic Network Slice Template (GST) and NEtwork Slice Type (NEST), which are used to quantify and qualitatively describe the requirements for an E2E NS [30]. The GST is a universal blueprint that contains a set of attributes used to characterize any NS. The NEST, on the other hand, is a filled-in version of the GST. Last but not least, the IETF has also been actively involved in developing high-level specifications for virtualization and network slicing in 5G and beyond mobile communication networks, covering the architectural frameworks, operations, maintenance, management, and orchestration, among many other aspects [31].

2) CONCEPTUAL AND ANALYTICAL MODELS FOR RAN SLICING IN THE NG-RAN ARCHITECTURE

In parallel to the SDOs and industry, a considerable amount of effort has been made in academia to study the extension of virtualization to the RAN architecture and the slicing of underlying resources. The overwhelming majority of available theoretical research focuses on *(i)* the slicing of radio resources [32]–[39], *(ii)* the allocation of virtual resources [40]–[42], *(iii)* the energy consumption of various types of RAN slices [43]–[45], and *(iv)* the application of AI techniques (notably ML-assisted algorithms) in the operations of RAN architecture [46], [47].

The aforementioned studies were the first to attempt to extend virtualization towards the edge of a cellular network and are considered the theoretical foundation for RAN slicing. However, the majority of these models are proposed in the context of cloud-RAN (C-RAN), fog-RAN (F-RAN), and heterogeneous-cloud RAN (H-CRAN) architectures, all of which were standardized by the 3GPP prior to Release 15. The NG-RAN is defined in Release 15 and is being developed further in Releases 16, 17, and 18. Therefore, the utilization of the above-cited models may not accomplish the desired performance objectives in the NG-RAN architecture.

Furthermore, while these studies addressed virtual resource allocation, they neglected the most critical stages that are prerequisite for RAN slicing, such as mapping VNFs and VLs onto the underlying infrastructure, designing the service function chain (SFC), and placing virtual machines (VMs) and VLs tailored to different types of RAN slices in the NG-RAN architecture. Last but not least, the aforementioned proposals are formulated based on traditional optimization models, which may not comply (due to their time-complexity) with E2E automation, zero-touch network and service management, and fully self-learning requirements of 5G and beyond mobile networks. Therefore, cutting-edge automatic data learning and synthesizing optimization tools are required to map, scale, configure, and allocate the virtual components of a RAN slice in a timely and resource-efficient manner.

Very recently, RAN slicing in the context of the NG-RAN architecture has sparked a lot of interest. We provide a comparative analysis of some of the recent works in Tab. 1. Among them, the authors of [10] experimentally evaluated the impact of virtualization on middlehaul latency in the NG-RAN architecture. The work of [11] was the first to propose a framework which dynamically provides RAN slice specific functional split options for enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-type communications (mMTC) types of services in the NG-RAN architecture. In [48], the authors proposed model-driven description files, which are used to automate the management and orchestration (among other aspects) of the virtual and physical components of the eMBB, URLLC, and mMTC types of RAN slices. However, in contrast to [11], the proposed model in [48] takes into account the static distribution of functionalities for three different types of RAN slices across the vCU, vDU, and RU.

Other studies have examined the deployment of VNFs and allocation of virtual compute resources [42], the analysis of various functional split options under energy and cost constraints [49], the mapping of RAN slices with a special emphasis on isolation and resource allocation [50], the placement of VNFs [51], the scheduling of VNFs for RAN slices [52], the impact of functional split on fronthaul delay in a RAN slice [53], and the sharing of VNFs across multiple network slice instances (NSIs) [54]. For the sake of clarity, the aforementioned theoretical models are compared

with respect to a number of attributes in Tab. 1 that can be clearly observed in the relevant columns.

3) MAPPING VNFs AND VLs ONTO LEGACY DATA CENTERS

In addition to the foregoing works, there exists a substantial body of literature focusing on the mapping of VNFs and VLs onto traditional cloud data centers and broadband networks. Among the most recent state-of-the-art solutions, for example, several key concepts and research challenges associated with the deployment of a SFC have been presented in [55]. A genetic algorithm is proposed for mapping the VNFs and SFCs onto optical networks in [17]. The authors of [56] used a column generation technique to address traffic routing and service scheduling. The placement of SFC in a federated multi-domain scenario has been examined in [57]. Reference [58] solved the admission control and SFC problems of the VNFs by employing relaxation, reformulation, and successive convex approximation methods. Concentrating on scheduling, [59] presented a genetic algorithm-assisted solution for scheduling the non-uniform incoming requests of VNFs. In [60], the instantiation of VNFs and the mapping of SFCs have been studied in wide area networks. Lastly, the authors of [61] proposed a joint algorithm for constructing the SFC and the mapping of VNFs onto cloud data centers in order to reduce the total bandwidth consumption of a network service.

The preceding studies and the references therein provided sufficient details on virtualization and softwarization, discovered various aspects of the mapping of VNFs and VLs onto the underlying NFVI-PoPs and transport networks, and demonstrated numerous advantages that virtualization brings to cloud and edge data centers. The models proposed in these works, however, have been evaluated in the context of traditional virtualized networks, neglecting the characteristics tailored to a wireless communication channel in both the upstream and downstream directions in the NG-RAN architecture. Therefore, deploying the above proposals may be inefficient in accommodating the divergent requirements of the VNFs and VLs of the bandwidth-devoring eMBB, latency-aware URLLC, and unlimited-things-centric mMTC types of RAN slices across a common 5G and beyond mobile communication infrastructure. Finally, the majority of the preceding studies are based on a single-objective optimization problem, solving either the mapping of VNFs or the mapping of VLs at a given time. However, due to the nature of the mapping problem, both VNFs and VLs are highly independent and may conflict with one another. Based on this, there is a need for a multi-objective and dynamic optimization solution, which must efficiently map the VNFs and VLs of a virtual network service onto the underlying NFVI-PoPs and transport networks at the same time.

B. THE PROBLEM

Despite widespread interest in virtualizing and intelligently managing the NG-RAN architecture, as well as the slicing of its underlying resources, the mapping process of the VNFs and VLs of a RAN slice – which is considered one of the most critical steps towards a virtualized, slicing-aware, and autonomous NG-RAN architecture – has not been fully addressed. This is due to the fact that (a) the development and deployment of virtualization, cloudification, and network slicing in the NG-RAN architecture are still in their early stages; (b) the state-of-the-art human-machine oriented interoperability between NFV-MANO, 3GPP-NSMS, and the underlying infrastructure is error-prone, slow, and cumbersome during the execution of the mapping process of the VNFs and VLs of a RAN slice; and (c) the mapping process of the virtual part of a RAN slice is inherently challenging because of the distinctive characteristics of the wireless communication channels (such as de[modulation], de[coding], de[multiplexing], de[ciphering], among others) and dynamic changes in the amount of virtual resources required by each VNF and VL.

Notwithstanding, the mapping process of the VNFs and VLs is also challenged by the density and behavior of the user equipments (UEs) of a RAN slice in a specific geographical region. Furthermore, the virtual resource allocation in the NG-RAN architecture, with the deployment of RAN slicing, has become more complicated due to new considerations such as the trade-off between utilization ratio and isolation level, the harmonization of inter-RAN and intra-RAN slice resource allocation algorithms, and the management of inter-RAN and intra-RAN slice priority. Therefore, the mapping of the virtual part of a RAN slice onto computing data centers and transport networks in the NG-RAN architecture is regarded as a valid research problem. Proposing an autonomous and intelligent solution to such a problem at an architectural level is the ultimate goal of this article.

C. OUR GOALS AND CONTRIBUTIONS

To contribute to filling the solution gap identified above, we propose an optimal architectural solution that is both autonomous and intelligent for mapping the virtual components of a RAN slice onto the underlying infrastructure in the NG-RAN architecture. To the extent of our knowledge, this is the first work addressing the mapping problem of a RAN slice in 5G and beyond mobile communication systems. The proposed architectural solution is distinct from the state-of-the-art alternative proposals discussed in the preceding sections and compared in Tab. 1 in terms of:

- *Application:* The proposed mapping solution is customized to the NG-RAN architecture and is specifically applicable to the mapping process of the virtual components of a single RAN slice.
- *Tool:* The proposed mapping solution employs an ML-assisted functioning architecture to execute the mapping process of a RAN slice, namely the ENI framework, which was recently introduced by the ETSI in order to integrate automation and intelligence into 5G and beyond mobile communication networks.

- *Architecture:* In contrast to the state-of-the-art alternative models, which assume solely the NFV-MANO framework, the proposed mapping solution in this article is executed on top of the unified architectural framework of the 3GPP-NSMS, NFV-MANO, ENI System, and the underlying infrastructure.
- *Scope:* The majority of the previous studies have concentrated on finding the optimal candidate NFVI-PoPs and transport links for mapping the VNFs and VLs of a network service. The focus of this study, however, goes beyond such a dilemma. This article investigates the components of the VNFs and the VLs of a RAN slice in greater detail, as well as their mapping onto their respective virtual environments in the underlying NFVI-PoPs and transport networks, aimed at optimizing a set of predefined optimization goals. These performance objectives could include decreasing the number of active physical components in an NFVI-PoP, minimizing the total bandwidth utilization of the PLs in the transport networks, reducing energy consumption both in the NFVI-PoP and transport networks, and so on.

In the light of foregoing goals, the main contributions of this article are thus:

- To extend the NGMN Alliance-defined functional architecture for network slicing towards the NG-RAN architecture, as well as to provide a comprehensive overview and critical analysis of the VNFs, PNFs, VLs, PLs, virtual and physical resources, and underlying compute and transport infrastructures that host the components of a RAN slice at the edge of a cellular network beyond the 5G communication system;
- To propose a unified architectural framework of the NFV-MANO FBs, the 3GPP-NSMS entities, and the underlying compute and networking infrastructures for managing and orchestrating the virtual and physical components of a RAN slice in the NG-RAN architecture;
- To adopt automation and intelligence into the management and orchestration of a RAN slice by integrating the ENI framework of the ETSI into the unified functioning architecture of the 3GPP-NSMS, NFV-MANO, and the underlying physical and virtual infrastructure in the NG-RAN architecture;
- To thoroughly define and model the mapping process of the components of the vCU and vDU, the VLs, and all the virtual aspects of a RAN slice at an architectural level, as well as to explore the internal domains and components of an NFVI-PoP and the underlying transport networks, based on a number of constraints that are critical for consideration in such a process;
- To propose a learning-assisted solution at an architectural level – leveraging ML techniques and AI tools in the joint framework of ENI, NFV-MANO, 3GPP-NSMS, and the underlying infrastructure – for mapping the components of the vCU and vDU, as well as the VLs, of a RAN slice onto their respective virtualized environments in the underlying edge data centers and transport networks in the NG-RAN architecture.

D. THE STRUCTURE OF THE ARTICLE

The rest of this article is structured in the following manner. We commence by extending the architectural framework of network slicing for RAN slices towards the NG-RAN architecture in Section II. In Section III, we delve deeper into the management and orchestration of the virtual components of a RAN slice, as well as integrate the ENI framework into the unified architecture of the NFV-MANO, 3GPP-NSMS, and underlying NFVI-PoPs aimed at bringing automation and intelligence to the management and orchestration of RAN slices. In Section IV, we define and model the problem of mapping the virtual components of a RAN slice onto the virtual resources in the underlying infrastructure. We then propose an ML-assisted solution, at an architectural level, for mapping the internal and external VLs, as well as the internal components of vCU and vDU, of a RAN slice onto the underlying NFVI-PoPs and transport networks in Section V. Lastly, Section VI summarizes the main conclusions of the article and provides future research directions in this area.

II. THE EXTENSION OF THE ARCHITECTURAL FRAMEWORK OF NETWORK SLICING TOWARDS THE NG-RAN ARCHITECTURE

In this section, we extend the architectural framework of network slicing, defined by the NGMN Alliance, towards the NG-RAN architecture in order to address the needs for an economically sustainable, intelligent, performance-aware, and energy-efficient RAN for 5G and beyond mobile communication networks [18], [62]. The overarching goal of this extension is to incorporate the three well-known categories of 5G communication services [63] – eMBB, mMTC, and URLLC – into the NG-RAN architecture. The proposed functioning framework, which is managed by the RAN management and orchestration plane, is shown in Fig. 1. It consists of three layers: the communication service layer, the network function layer, and the infrastructure layer. These three layers, as well as their management and orchestration plane, are described in the following subsections, respectively.

A. THE COMMUNICATION SERVICE LAYER

This layer defines the types, behavior, and characteristics of the communication services provided to the tenants or MVNOs by the RAN slices in the NG-RAN architecture. The characteristics of the communication services, the isolation and allocation of required resources, the life cycle management, and other technical requirements are defined in a deployment descriptor file called the RAN network slice subnet template (NSST) [64], as shown in Fig. 1. The RAN NSST is identified based on a network slice

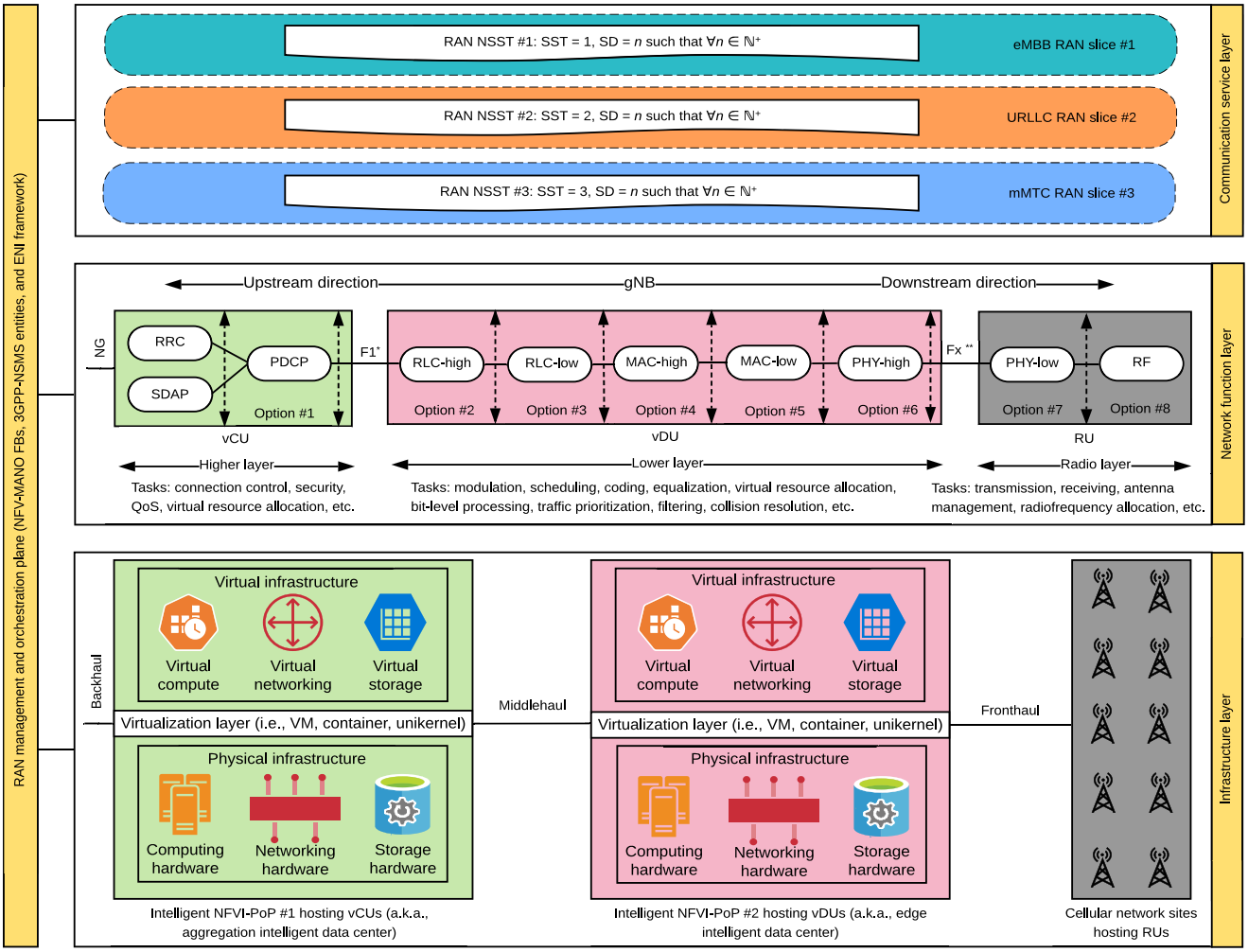

FIGURE 1. **The proposed architectural framework for RAN slicing in the NG-RAN architecture. Do note that (∗) F1 denotes a standard interface between the vCU and vDU, whereas (∗∗) Fx is used as a generic notation for functional split between the vDU and RU.**

selection assistance information (NSSAI), which is specified by the 3GPP in [65]. In order to assist the 5G network in selecting a RAN slice, the UE must send NSSAI in a signaling message towards the NG-RAN architecture. According to 3GPP specifications [65], the NSSAI consists of a set of eight single NSSAIs (S-NSSAIs), each of which uniquely identifies a RAN slice. This means that a UE can potentially be served by a maximum of eight RAN slices at the same time, each with its own customized RAN NSST.

Each S-NSSAI is a 32-bit parameter that is comprised of two components (also called fields): the slice/service type (SST) and the slice differentiator (SD) [22].

- The SST is a mandatory 8-bit component that defines the type of communication services, features, and behavior of a RAN slice. The SST field contains a specific value that can be either standardized or non-standardized (operator-specific). The standardized SST range is from 0 to 127, while the operator-specific SST range includes values from 128 to 255 [65]. For eMBB the SST $= 1$, for URLLC the SST $= 2$, and for mMTC the SST $= 3$. In addition to the three SSTs mentioned previously, two further SSTs have been standardized by the 3GPP. They are SST $= 4$ for vehicle-to-everything (V2X) and SST $= 5$ for high-performance machine-type communications (HMTC). For the sake of simplicity, we consider only the SSTs that belong to the eMBB, URLLC, and mMTC types of RAN slices in this article. We assume that the remaining two SSTs are managed in the same manner by the proposed functioning architecture.
- The SD is an optional 24-bit component, complementing the SST. It is used to identify (or differentiate) multiple homogeneous RAN slices offered to the same or different tenant(s). This differentiation could be made on the basis of features, applications, network coverage zones, and priority, among other criteria. The SD field is also provided with a value, ranging from 1 to $n \in \mathbb{N}^+$. For example, a tenant (or MVNO) may request two eMBB RAN slices and one URLLC RAN slice from

a network operator. In this case, the first eMBB RAN slice is identified with SST = 1 and SD = 1, the second eMBB RAN slice is identified with SST = 1 and SD = 2, while the URLLC RAN slice is identified with SST = 2 and SD = 1.

B. THE NETWORK FUNCTION LAYER

The network function layer defines the radio processing functionalities, which are implemented as VNFs or PNFs and constitute the eMBB, URLLC, and mMTC types of RAN slices (see Fig. 1). These functionalities are distributed across the NG-RAN architecture in a flexible manner. The NG-RAN architecture is composed of a set of $n \in \mathbb{N}^+$ gNBs [66]. The gNB incorporates three functional modules: the vCU, the vDU, and the RU [67]. These components could be deployed in various geographical locations based on the network operator's constraints and the tenant's requirements [23]. Each gNB consists of at least one vCU. Depending on network topology and the density of UEs in a specific geographical region, each vCU may correspond to multiple vDUs, and each vDU may correspond to multiple RUs [66]. For the sake of simplicity, we assume a vCU, a vDU, and an RU belong to a gNB in this article, as shown in Fig. 1.

These three logical functional modules of a gNB may be provided by the same vendor or by multiple vendors. Thus, the NG-RAN architecture must support seamless interoperability between the hardware and software components from multiple vendors while participating in the creation of a RAN slice [68]. Such an openness and interoperability in the NG-RAN architecture improves several features, including network malleability, programmability, intelligence, and security [69], [70]. It also reduces total network expenditure, extends virtualization to the extreme edge, strengthens the management and orchestration framework, brings versatility to the market, and improves underlying resource allocation. To facilitate open implementation of the components of a gNB, the O-RAN Alliance and the Telecom Infra Project (TIP) have been developing specifications over the last several years [15]. These components of a gNB are denoted as Open-CU, Open-DU, and Open-RU in the O-RAN Alliance's specifications [28], [69].

The radio processing functionalities of a gNB are dynamically distributed into eight split options (option #1 to option #8) across the vCU, vDU, and RU [67]. The overarching goal of such a functional distribution is to enable virtualization and softwarization of eMBB, URLLC, and mMTC types of RAN slices over a common NG-RAN infrastructure, as well as to efficiently allocate their required virtual and physical resources [71]. The distribution of the gNB functionalities among its components, also called the functional split, determines which functions are processed in the vCU, which are located in the vDU, and which are moved towards the RU in both downstream and upstream directions.

The 3GPP initially considered option #2 as a functional split between the vCU and vDU and option #7 or option #8 between the vDU and RU [53]. However, other SDOs, such as Small Cell Forum (SCF), evolved common public radio interface (eCPRI) Cooperation Group, and O-RAN Alliance moved further in identifying new options for functional-splits in the NG-RAN architecture while taking user density, bandwidth, and latency requirements of RAN slices into account. These recently introduced functional split options are: (a) option #1 between the vCU and vDU; and (b) option #6, option #7-1, option #7-2a, option #7-2, and option #7-3 between the vDU and RU. These split options are analyzed and compared in [20], [21], [27], [72], [73]. In this article, our goal is not to discuss the distribution of gNB functionalities, their respective architectures, and their implications on the performance of RAN slices. For the sake of simplicity, we thus only assume: (*i*) the three aforementioned components of a gNB that accommodate these functionalities physically and virtually; (*ii*) option #2 for the vCU and vDU split, which is specified over a well-defined logical interface (F1); (*iii*) option #7 for the vDU and RU split over a standard interface (Fx); and (*iv*) both F1 and Fx have the capability of dynamically supporting different requirements (such as the data rate, latency constraints, user density, high mobility, and others) of eMBB, URLLC, and mMTC types of RAN slices in the NG-RAN architecture.

The functionalities of the vCU and vDU are either already virtualized or in the process of being virtualized. Those that have not yet been virtualized require additional research in terms of their softwarization and cloudification. However, there is no doubt that they will be fully virtualized in the near future or long term [70]. Based on this, the functionalities that are accommodated by the vCU (also known as the higher layer functionalities) and vDU (also known as the lower layer functionalities) are assumed to be fully virtualized and could be instantiated as VNFs in an aggregation data center and an edge data center, respectively. The higher layer consist of three major functionalities: the radio resource control (RRC), the service data adaptation protocol (SDAP), and the packet data convergence protocol (PDCP). The lower layer consist of five major functionalities: the radio link control (RLC)-high, the RLC-low, the medium access control (MAC)-high, the MAC-low, and the physical (PHY)-high. To gain a better understanding of the aforementioned functionalities, we provide a brief discussion on their definitions and objectives in a gNB in Tab. 2. The interested readers are also suggested to refer to the relevant publications, such as [43], [72], [73], as well as the references therein, for a more in-depth understanding. These eight virtual functionalities of the vCU and vDU – as well as the VLs that connect them – running in aggregation and edge data centers require virtual resources such as virtual compute, networking, and storage [70], [74].

The functionalities of the RU, which are also known as the radio layer functionalities, include the PHY-low and radio frequency (RF). These two functionalities are assumed to be the PNFs of a gNB. Both of these physical functionalities and the PLs connecting them are implemented on top of customized equipment and physical resources such as antennas,

TABLE 2. The radio processing functionalities that comprise a RAN slice in both upstream and downstream directions, as well as the functional split options of a gNB [75]. It should be noted that these eight possible split options are proposed by the 3GPP as part of a study item of the NG-RAN architecture in Release 15 and have further expanded in Releases 16 to 18.

Split option	RAN slice functionality	Objective in a RAN slice
Option #1	RRC/SDAP – PDCP split	This layer includes a subset of the gNB functionalities (specifically, the SDAP, RRC, and PDCP), which are related to the vCU [75]. The SDAP sublayer is responsible for handling the quality of service (QoS) flow across the RAN slice. This is accomplished by mapping a protocol data unit (PDU) session (with a defined level of QoS) received from a CN NSS to its corresponding data radio bearer (DRB) of a RAN slice. Furthermore, the SDAP also marks the downlink and uplink packets with QoS flow IDs (QFIs) to ensure that they receive proper forwarding treatment in a RAN slice. The RRC sublayer performs a number of functions on the control plane, such as the establishment and release of connections and DRBs, paging notification, security and mobility procedures, QoS management, and others. It also allows radio resource management strategies to be implemented and configures the data and control planes by means of signalling functions in a RAN slice. The PDCP sublayer applies a number of control and user plane functions on DRBs, including ciphering and integrity protection, packet duplication, transfer of user and control plane data, header compression of internet protocol (IP) packets, and others in a RAN slice. This layer connects the vCU to the 5G core network (5GC) over the next-generation (NG) logical interface (see Fig. 1). The data rate requirement over the NG interface could be up to hundreds of Gbps, the latency requirement is in the order of 40 ms, and the distance requirement between the vCU and 5GC could be up to 200 km.
Option #2	PDCP – RLC-high split	This layer splits the vCU and vDU via the F1 logical interface, as shown in Fig. 1. The data rate required over the F1 interface is approximately 100 Gbps, the latency requirement is in the ms range, and the distance requirement between the vCU and vDU could be up to 80 km. The gNB functionalities from option #2 to option #6 are assumed to be related to the vDU in this article. This layer is internally divided into two layers: the RLC-high and the RLC-low. The former is responsible for the automatic repeat request (ARQ) function as well as other receiving entities such as unacknowledged mode (UM), transparent mode (TM), and the status reports related to uplink transmission in a RAN slice [71], [75].
Option #3	RLC-high – RLC-low split	The RLC-low is the second functionality that is hosted by the vDU. It is composed of segmentation functions as well as other transmitting entities such as TM, UM, acknowledged mode (AM) routing, and the status reports related to downlink transmission in a RAN slice [71].
Option #4	RLC-low – MAC-high split	The MAC-high layer is responsible for functions such as centralized scheduling, encapsulating frames for transmission over physical medium and vice versa, determining the channel access methods for transmission, generating frame check sequences to protect them against errors, collision resolution and initiating re-transmission, multiplexing and demultiplexing of PDUs, traffic prioritization using QFIs, and coordinating inter-cell interference in a RAN slice [20].
Option #5	MAC-high – MAC-low split	The MAC-low layer hosts functions that (a) optimize the performance of a RAN slice, such as hybrid ARQ (HARQ), (b) process the scheduling-related information, (c) comprise a number of radio resource management (RRM) procedures, and (d) map logical channels to transport channels and vice versa. It also measures and estimates the activities and configured operations of the UEs of a RAN slice and subsequently reports them to the MAC-high layer on a periodic basis [20], [71].
Option #6	MAC-low – PHY-high split	The PHY-high is the last layer which is assumed to be hosted by the vDU in this article, as shown in Fig. 1. It is responsible for performing a number of functions in a RAN slice, such as modulation and demodulation, coding and decoding, load balancing, precoding, and possibly pre-filtering in both downlink and uplink directions [71], [75].
Option #7	PHY-high – PHY-low split	This layer splits the vDU and RU by means of the Fx logical interface (see Fig. 1). The data rate requirement over the Fx interface is 20 Gb, the latency requirement is in the ms range, and the distance requirement between the vDU and RU could be up to 20 km. It is responsible for aspects such as resource block allocation, beamforming, antenna configuration, and other functions that are not executed in the PHY-high layer of a RAN slice [71].
Option #8	PHY-low – RF split	This layer provides wireless connectivity to the UEs via a new radio (NR) protocol stack that employs a specific range of RF. The RF blocks are configured by the PHY-low layer and are dynamically allocated to the UEs of a RAN slice using an antenna in this layer [20].

radio frequency spectrum, Ethernet cables, optical fiber, and other dedicated hardware installed on Macro, Micro, Pico, and Nano cellular network infrastructure sites [70].

In order to improve resource utilization, enhance energy efficiency, decrease CAPEX and OPEX, and increase the number of served tenants over a common network infrastructure, it is expected that the aggregation data center, edge data center, and cellular site should have the capability to host, configure, and allocate the required resources of the vCU, vDU, and RU functionalities of various eMBB, URLLC, and mMTC types of RAN slices concurrently. This specifically means that a gNB could provide $n \in \mathbb{N}^+$ RAN slices of various types at the same time. Based on this, in this article, we assume that the virtual and physical resources of the gNB's components are dynamically shared among the three aforementioned types of RAN slices over the NG-RAN architecture.

C. THE INFRASTRUCTURE LAYER

The infrastructure layer configures and allocates the virtual and physical resources of a RAN slice in virtualized and physical network sites, respectively (see Fig. 1). The physical sites are the cellular network infrastructure sites, which are classified as Macro, Micro, Pico, and Nano cellular sites. They are distributed uniformly or non-uniformly over a specific geographical region based on network topology, user density, and other parameters associated with network design [76]. Each cellular network site consists of an antenna, a tower, and communication control equipment, which are used to deploy the physical functionalities of the RU (namely

the PHY-low and RF layers) for the eMBB, URLLC, and mMTC types of RAN slices. The virtualized network sites are the cloud sites located at the cellular network's edge, namely the aggregation data center and edge data center. Each data center is also called the NFVI-PoP. The NFVI-PoP consists of general-purpose hardware and software components that are used to host, manage, and execute the virtual functionalities of the vCU (namely the RRC, SDAP, and PDCP layers) and vDU (namely the RLC-high, RLC-low, MAC-high, MAC-low, and PHY-high layers) of the three types of RAN slices [77]. The VNFs in NFVI-PoP #1 (namely the vCUs) and NFVI-PoP #2 (namely the vDUs) are usually hosted in different geographical locations, as shown in Fig. 1. Depending on the deployment scenario, the distance between NFVI-PoP #1 and NFVI-PoP #2, as well as between NFVI-PoP #2 and cellular network sites, could be up to tens of kilometers [72]. Therefore, the vCUs and vDUs hosted in NFVI-PoP #1 and NFVI-PoP #2 are interconnected using a reliable F1 middlehaul transport link. The vDUs in NFVI-PoP #2 are connected to the RUs at the cellular sites via a reliable Fx fronthaul transport link [73] (see Fig. 1).

Focusing on the virtualized network sites of the infrastructure layer in this article, both NFVI-PoP #1 and NFVI-PoP #2 are used to host a large number of vCUs and vDUs of various types of RAN slices that have complex and potentially conflicting demands and requirements, respectively [78]. The state-of-the-art NFVI-PoPs manage, orchestrate, and allocate the virtual resources of the vCUs and vDUs using conventional methods (or traditional algorithms). However, the performance of such mechanisms is insufficient to accommodate the requirements of a large number of vCUs and vDUs in a shared and heterogeneous NG-RAN architecture with constrained virtual resources [79], [80]. To adjust various types of services of RAN slices based on dynamic changes in the tenant's domain, business objectives, and environmental conditions, we extend the utilization of AI techniques – specifically the ML-assisted algorithms such as supervised, unsupervised, reinforcement learning (RL), and deep learning (DL) – to the NFVI-PoP #1 and NFVI-PoP #2. The grand objective of such an intelligentization of the NFVI-PoPs is to automate the management, orchestration, and allocation of the virtual resources of the vCUs and vDUs. In this article, we refer to such AI-empowered NFVI-PoPs as Intelligent PoPs (I-PoPs). Each I-PoP employs ML-assisted algorithms in order to optimize the performance of various types of RAN slices, resulting in the overall operational enhancement of the cellular network.

In addition to the above, the state-of-the-art NFVI-PoPs are based on human-machine interaction models, which are slow, expensive, error-prone, and cumbersome [78], [81]. For example, each NFVI-PoP in a 5G mobile communication system is expected to be a complex network infrastructure comprised of software and hardware components from multiple vendors. It must support a large number of RAN slices with customized services that can be dynamically scaled up and down or scaled in and out. The conventional human-machine interaction models – between vendors' components and the operator's management system, as well as between the tenants and operator – in such a heterogeneous NFVI-PoP are not adaptable to dynamic changes in vCUs and vDUs of various types of RAN slices. These existing research challenges in traditional NFVI-PoPs may result in extremely high CAPEX and OPEX for resource deployment and management in a cellular network. To minimize CAPEX, OPEX, and human-dependent decision making processes, operators need the ability to automate and intelligentize several aspects, such as the deployment, configuration, optimization, monitoring, management, and orchestration of the resources and operations in I-PoPs. Therefore, the utilization of cutting-edge ML algorithms – specifically, RL, DL, and federated learning (FL) – in I-PoPs has profound implications on reducing CAPEX and OPEX, decreasing energy consumption, and lowering management and network complexity in the NG-RAN architecture [82].

The bottom side of Fig. 1 also depicts that each I-PoP is composed of physical infrastructure, virtualization layer, and virtual infrastructure [83].

The physical infrastructure is characterized by compute, storage, and networking hardware such as nodes, devices, and links, respectively. They provide compute resources, storage capacity, and internal/external connectivity for the vCU and vDU [83], [84]. The compute node is a general-purpose computing hardware unit that is managed by its internal instruction set and realized in a single or multiple central processing unit (CPU) servers such as tower servers, blade servers, rack servers, and others. The storage device is capable of temporarily or permanently storing a large amount of data (or information) related to the vCU, vDU, and the internal and external VLs of a RAN slice. It may be used in the form of direct attached storage (DAS), network attached storage (NAS), and storage area network (SAN). The networking links facilitate communication channels between the compute nodes of the I-PoPs and other elements (including the storage device) in the form of layer-2/layer-3 (L2/L3) or bare-metal switches.

The virtualization layer is a software platform that is placed between the physical infrastructure and the virtual infrastructure in an I-PoP (see Fig. 1). In this layer, the underlying physical resources are abstracted into several isolated virtual environments of the virtual compute, networking, and storage resources [83]. The virtual environments of the virtual networking and storage resources are the virtual networking (VN) and the virtual memory, respectively [84]. The notion of VN is frequently used to refer to a virtual environment that enables virtual communication between two or more VNFs. We have noticed that several articles referred to the terms VN and VL interchangeably. To the best of our knowledge, the former is abstracted directly from the underlying physical resource in order to host the latter (which is a logical component of a RAN slice). The VN connects the vCU, vDU, and their respective functionalities virtually to form a complete RAN slice. The virtual memory virtually

stores data associated with the vCU, vDU, and their respective VLs throughout the life cycle of a RAN slice. The VN (or VL) and virtual memory are covered in more detail later in the article.

The virtual environments for the virtual compute resources are VMs, containers, and unikernels [83]. The VM creates an isolated virtual environment identical to that of a physical compute node for the purpose of hosting VNFs. The container includes only the necessary elements for hosting VNFs. The unikernel is an ultra-lightweight, single-purpose virtual environment that runs only a single VNF or a single application of a VNF [83]. The VM and unikernel are achieved using hypervisor (which will be discussed later), such as Hyper-V, ESXi, Xen, KVN, etc. The container is abstracted by container technology such as FlowN, Docker, LXC, etc. In each scenario, the performance of the virtual compute environment must be functionally equivalent to that of the physical compute environment. Each of the aforementioned virtual environments has unique characteristics and is suitable for specific virtual compute scenarios. Interested readers may refer to [83] for more detailed information.

Furthermore, to provide underlying virtual compute resources with E2E automation and strong isolation, the use of container in virtual machine (CVM) has recently emerged as a virtual environment [85] that is capable of hosting the virtual components of a VNF. The grand objective of this approach is to reap the benefits of both containers and VMs concurrently. For example, the VM guarantees the isolation of virtual resources, while the container simplifies the execution of the VNF that requires these resources. The CVMs can be manifested by the hypervisor as well. The hypervisor (*i*) manages the CVMs (and VMs) and the allocation and reallocation of their respective virtual resources; (*ii*) provides isolation among the virtual resources of CVMs (and VMs); (*iii*) maps the relationship between physical and virtual resources allocated to the CVMs (and VMs); (*v*) schedules the virtual resources of the CVMs (and VMs); and (*iv*) emulates the hardware components in such a way that would make the CVMs (and VMs) appear to be running on real dedicated devices through specific application programming interfaces (APIs) [83].

Given the required isolation of the virtual compute resources, which is essential for RAN slicing as well as supporting multi-tenancy by underlying I-PoPs, none of the aforementioned virtual environments can standalone efficiently fulfill the requirements of the eMBB, URLLC, and mMTC types of RAN slices due to their disparate resource demands, service characteristics, and performance metrics. Therefore, there is a need for additional research efforts to specify which type of virtual environment is appropriate for which type of RAN slice in the NG-RAN architecture. For the sake of simplicity, we assume VM as a virtual environment that is used to host the virtual compute resource of the three types of RAN slices discussed in this article.

The virtual infrastructure is the third layer of an I-PoP (see Fig. 1), which is composed of virtual compute, networking, and storage resources. As previously stated, the virtual resources are abstracted, using the hypervisor, from the underlying physical resources such as compute nodes, PLs, and storage devices, respectively. These virtual resources are discussed in greater detail in the following.

The virtual compute resources are created by virtualizing the physical processing nodes or elements (specifically, the CPU) using a certain compute virtualization technology such as software-defined compute (SDC), vCenter Converter, or Libvirt, among others. These technologies also move the compute resources to a resource pool, which are then assigned based on a dynamic compute resource allocation algorithm to the VMs. The VMs and the compute resources therein are then used to host the vCU and vDU of a RAN slice. From the resource allocation perspective, the amount of compute resources assigned to a VM is typically quantified in terms of the number of virtual CPUs and clock speeds.

The virtual networking resources are decoupled from the underlying PL resources (such as the Ethernet cables, optical fibers, L2/L3 or bare-metal switches, etc.) using a virtual software application called the virtual switch (vSwitch). These virtual networking resources connect VMs, containers, unikernels, CVMs, virtual servers, and other elements of the virtual infrastructure layer within the same or across different I-PoPs. The VNs that connect the VMs belong to the vCU and vDU of a RAN slice are typically measured by the number of total allocated VLs and the bandwidth of each of them in kilo bits per second (Kbps).

The virtual storage resources are abstracted from the underlying hardware data storage resources (such as memory blocks and storage media) in the form of DAS, NAS, or SAN using a certain storage virtualization technology such as software-defined storage (SDS) and made available in a virtual storage resource pool. Based on the requirements of the VMs and VLs, the storage resources are dynamically allocated to the components of a RAN slice in order to store their data in a virtual storage resource pool, either temporarily or permanently. The virtual storage resources of the VM and VL are usually measured in the kilobyte (KB).

D. THE MANAGEMENT AND ORCHESTRATION PLANE

The management and orchestration plane is playing an essential role in the assurance of the efficient utilization of the underlying resources and the tight integration of the aforementioned three layers – the communication service layer, network function layer, and infrastructure layer – while also meeting the performance, operational, and functional requirements of the offered services of the proposed RAN slicing framework for the NG-RAN architecture. The management and orchestration plane is depicted in Fig. 1 as being made up of the 3GPP-NSMS, NFV-MANO, and ENI System. The 3GPP-NSMS (described in the following section) is responsible, among other tasks, for all operations relating to the physical resources, connectivities, and services of RAN slicing in the NG-RAN architecture. On the contrary, the

NFV-MANO (explained further below) is in charge of the virtual resources, connectivities, components, and operations associated with RAN slicing. The unified integration of the NFV-MANO and 3GPP-NSMS is thus critical for managing and orchestrating both the physical and virtual components of the three types of RAN slices in the NG-RAN architecture.

The state-of-the-art integration of the NFV-MANO and 3GPP-NSMS is based on human-machine-type interaction models that employs standard interfaces. For more effective control and orchestration of the resources and services, operators demand an automated arrangement and coordination of both management systems in the NG-RAN architecture. Inherent intelligence and implicitly autonomous control of all services, layers, and underlying resources are thus required at the edge of a cellular network to meet the zero-touch management, orchestration, and operation requirements of the RAN slices. In response to the industry's demand for such an AI/ML-driven intelligent mobile network, the ETSI established the ENI ISG in February 2017 [86]. As of this writing, the ENI ISG has 44 members and 21 participants representing academia, industry, vendors, and research institutions. The ENI System does not necessarily interfere with (or alter) the management, orchestration, resource allocation, and other operations covered by the 3GPP-NSMS and NFV-MANO specifications. Nonetheless, its overarching goals are to optimize those operations and automate complex human-dependent decision-making processes through the use of cutting-edge ML-assisted tools and methods, as well as context-aware and metadata-driven policies [86].

The purpose of the interoperability and integration of the ENI System into NFV-MANO and 3GPP-NSMS is to automate the management and orchestration, among other aspects, of RAN slices with the goal of improving the performance of the NG-RAN architecture. On the one hand, the NFV-MANO and 3GPP-NSMS use a variety of tools and data ingestion formats to interact with each other. The ENI System, on the other hand, employs its own customized modules, each with its own exposed APIs and varying degrees of capability. As a result, it was extremely important for the ENI ISG to define an architecture that converts the input of the NFV-MANO and 3GPP-NSMS to a format understandable by the ENI System and vice versa [81]. In order to achieve this goal, the ENI ISG has proposed a modularized system architecture in [26]. We extend the ENI proposed framework towards the NG-RAN architecture and integrate it into the 3GPP-NSMS and NFV-MANO in order to automate the management and orchestration of services and resources of various types of RAN slices in 5G and beyond mobile networks. The following section provides a detailed description of such an extended architectural framework in the context of the NG-RAN architecture.

III. THE MANAGEMENT AND ORCHESTRATION OF RAN SLICES IN THE NG-RAN ARCHITECTURE

Following the extension of the network slicing architectural framework towards the NG-RAN architecture, this section provides a deeper insight into the management and orchestration plane, which is responsible for managing the VNFs and VLs of a RAN slice and orchestrating their respective virtual resources in underlying I-PoPs and transport networks. To accomplish this, we will first discuss a joint framework of the NFV-MANO and 3GPP-NSMS used for the management and orchestration of RAN slices in the NG-RAN architecture. Following that, we will integrate the ENI System into the 3GPP-NSMS and NFV-MANO in order to bring intelligence and automation to the edge of a 5G cellular network through the use of ML-assisted algorithms. The ultimate goals of the proposed unified framework are to: (a) enable the NG-RAN architecture to be more reliable and maintainable; and (b) provide context-aware and ML-assisted services in order to meet the business objectives and technical requirements of the tenants and MVNOs.

A. THE 3GPP-NSMS FOR RAN SLICES

Fig. 2 depicts the functioning architecture of the 3GPP-NSMS as part of its joint framework with the NFV-MANO. At the start, the communication service management function (CSMF) receives requirements related to the communication service from a tenant or MVNO via a set of APIs, translates them to the NS-related technical requirements, and delivers them to the network slice management function (NSMF) in a deployment descriptor file called the network slice template (NST). Among other attributes, the NST defines the required physical and virtual resources, the interconnection and configuration of these resources, and the life cycle management of an E2E NS.

To this end, the 3GPP has defined the eMBB, URLLC, and mMTC types of NSTs. The NSMF employs the NST to manage its specific type of NS throughout its life cycle and allocate its customized physical and virtual resources in isolation from those of other NSs over a shared infrastructure [87]. The 3GPP has divided the NST into CN NSST and RAN NSST [1]. Both NSSTs, among other attributes, define the required physical and virtual resources for a CN NSS and a RAN slice in 5G core network (5GC) and NG-RAN, respectively. The transport network NSS, which connects the CN NSS and the RAN slice, is also instantiated based on a transport network NSST. The transport network NSST defines, among other attributes, the required physical and virtual networking resources for the communication channels of an E2E NS.

To manage an E2E NS effectively and orchestrate its associated resources efficiently, the NSMF divides the requirements related to an NS into CN NSS, transport network NSS, and RAN NSS related requirements [87]. Following that, the physical and virtual resources of each NSS, as well as all the events that occur during its lifetime, are independently managed by a 3GPP proposed management entity called the network slice subnet management function (NSSMF) [87]. This is also depicted in Fig. 2, where the NSMF delegates the requirements of the NSSs to their respective NSSMFs. Among them, the NG-RAN NSSMF

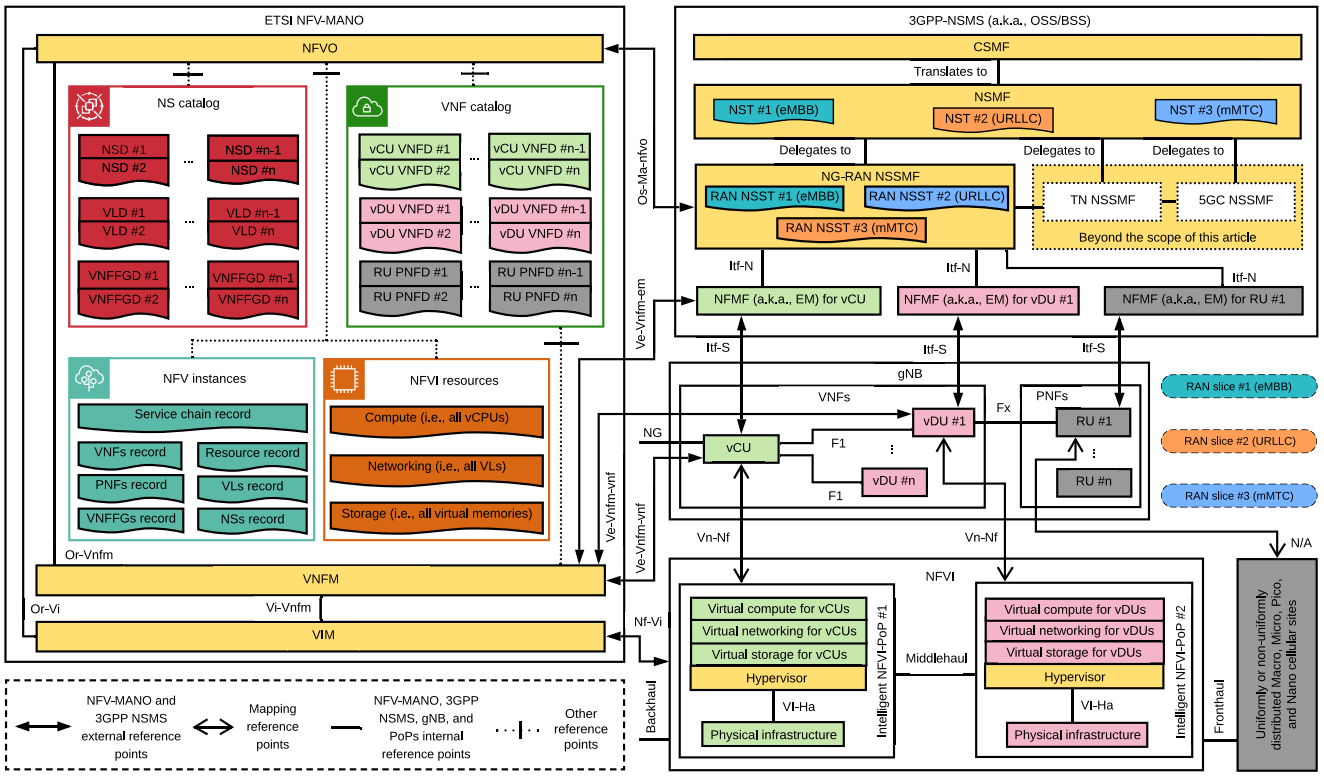

FIGURE 2. **The management and orchestration of RAN slices in the NG-RAN architecture, using the 3GPP-NSMS and NFV-MANO joint framework. It should be noted that the VNFM is directly connected to each of the VNFs (namely the vCU and vDUs) of a gNB, their corresponding NFMFs, and the NFMFs of PNFs (namely the RUs). Due to space limitations in the figure, we show its connection with vCU, vDU #1, and NFMF for vCU. The rest of the NFMFs and VNFs are connected to VNFM in the same vein. Do also note that the detailed descriptions of the 3GPP-NSMS entities and NFV-MANO FBs, as well as their interactions over standard interfaces, are intentionally left abstract in this article. Interested readers are advised to refer to the relevant publications. In this figure, NS catalog = network service catalog, NSs = network services, and TN = transport network.**

is responsible for the allocation, deployment, fault management, and performance monitoring of virtual and physical resources of eMBB, URLLC, and mMTC types of RAN slices in a specific geographical region covered by at least one gNB belonging to that NG-RAN architecture.

The 3GPP further hierarchically divides the tasks of the NG-RAN NSSMF into a number of lower level management entities [13], referred to as network function management functions (NFMFs). The NFMF, also known as the element manager (EM), is responsible for managing the fault, configuration, accounting, performance, and security (FCAPS) of each type of component of a gNB (i.e., vCU, vDUs, and RUs) or each of the components of a gNB individually. For example, there could be three NFMFs to manage the FCAPS of a gNB, each of which is responsible for the vCU, vDUs, and RUs, respectively; or a number of NFMFs could be assigned to manage the FCAPS of each vCU, vDU, and RU in each gNB. The NFMF is controlled by the NG-RAN NSSMF in in both of these scenarios.

B. THE NFV-MANO FOR RAN SLICES

In previous sections, we assumed that the vCU and vDU are the VNFs of a RAN slice. The VLs that connect the internal components of these VNFs are said to be the internal VLs. The connection of the vCU and vDU among each other and with the 5GC and RU are established by external VLs. These VNFs and VLs of a RAN slice require virtual compute, networking, and storage resources. The allocation of virtual resources, the life cycle management of the VNFs and VLs, and the connectivity of the VNFs to PNFs are not in the scope of the 3GPP-NSMS. The ETSI NFV-MANO is in charge of carrying out these tasks. The NFV-MANO is made up of descriptors, repositories, internal and external reference points, and FBs. The FBs of the NFV-MANO are connected to the entities of the 3GPP-NSMS (which is also known as the OSS/BSS) via standardized interfaces aimed at establishing a unified framework for the management and orchestration of the virtualized and physical parts of a RAN slice in the NG-RAN architecture. Such a unified architectural framework is depicted in Fig. 2.

The components of the virtualized part of a RAN slice (which are managed by the NFV-MANO) are the RAN functions (i.e., vCUs and vDUs), external and internal VLs, the connection points of the VLs, the interconnection of the RAN VNFs to PNFs, and the virtual network function forwarding graphs (VNFFGs). The vCU, vDU, and their respective internal functionalities are connected by the VLs. The VL also connects the VNFs of a RAN slice to the RU. The ETSI proposed the VNFFG to describe the topology of the virtualized part of a RAN slice by specifying how

the connection points of the VLs are used to connect the vCU, vDU, and RU. These virtual components are present in all types of RAN slices. The NFV-MANO leverages customized machine-processable descriptor files for each of the given components to manage their life cycles and allocate their corresponding virtual resources with full automation, agility, and effective scaling.

The descriptors are the deployment templates that define the operational, performance, functional, and policy requirements, as well as the required resources, of each of the virtual components of a RAN slice [77]. They are written in data modelling languages such as YAML, TOSCA, YANG, or others [88]. The information contained in these model-driven descriptors enables the NFV-MANO to manage the vCU, vDU, and their associated virtual resources with full automation and a great level of control throughout their life time. These descriptors – which may have been previously on-boarded or newly generated prior to being on-boarded on the network function virtualization orchestrator (NFVO) – are generally related to *(i)* the NFs, namely virtual network function descriptors (VNFDs) for vCUs and vDUs and physical network function descriptors (PNFDs) for RUs; *(ii)* the network services, specifically the network service descriptors (NSDs); *(iii)* the VLs, namely the virtual link descriptors (VLDs); and *(iv)* the topology, precisely the virtual network function forwarding graph descriptors (VNFFGDs). Among them, the descriptors related to the virtual resources (which are integrated into the VNFDs of the vCU and vDU) and topology (specifically, the VNFFGDs) are discussed in Section IV. As a result, we will refrain from providing additional information on the remaining NFV-MANO and 3GPP-NSMS description models. However, interested readers are advised to refer to [30] for more details on the 3GPP-NSMS related templates, as well as to [25], [48], and the references therein for more insights into the NFV-MANO related descriptors.

The description models and other necessary (profile and behavioral) information related to the virtualized parts of eMBB, URLLC, and mMTC types of RAN slices are predefined and stored in four repositories: the network service catalog, the VNF catalog, the NFV instances repository, and the network function virtualization infrastructure (NFVI) resources repository. Fig. 2 illustrates that these four repositories, respectively, contain the (a) NSDs, VLDs, and VNFFGDs; (b) VNFDs for vCUs, VNFDs for vDUs, and PNFD for RUs; (c) information about all VNF instances and network service instances during their life cycles; and (d) information about the NFVI compute, storage, and networking resources (including their status such as available, allocated, reserved, utilized, and wasted) of all running components of RAN slices.

The NFV-MANO is composed of three FBs that use description models in order to manage underlying resources efficiently, operate multiple RAN slices across a shared NG-RAN architecture dynamically, and orchestrate the virtualized components presented in all types of RAN slices effectively. These FBs are the NFVO, virtual network function manager (VNFM), and virtualized infrastructure manager (VIM).

- The NFVO on-boards and creates network service instances, VNFFG instances, and VNF instances associated with eMBB, URLLC, and mMTC types of RAN slices using their predefined description models. The NFVO also validates and authorizes virtual resource requests for a RAN slice. Furthermore, it communicates with the NG-RAN NSSMF in order to jointly instantiate the virtualized part of the requested RAN slice.
- The VNFM manages the life cycle of the VNF instances (specifically, the vCU and vDU) and VLs of a RAN slice. It also governs the overall performance and fault events associated with the vCU, vDU, and VLs throughout the lifetime of the RAN slice.
- The VIM allocates, controls, and manages the virtual compute, storage, and networking resources required by each of the vCUs and vDUs of all types of RAN slices within their respective I-PoPs.

As illustrated in Fig. 2, the NFV-MANO FBs, the 3GPP-NSMS entities, and the VNFs and PNFs of a gNB are interconnected using standard reference points defined by 3GPP and ETSI. Employing such an interoperable functioning framework, the physical and virtualized components of a RAN slice deployed over I-PoPs and cellular network infrastructure can be effectively managed, and their respective physical and virtual resources can be efficiently allocated.

C. THE UNDERLYING I-POPS FOR HOSTING THE RAN SLICES

We argued in the previous sections that the NFVI-PoPs host the virtual resources of vCU and vDU, which are critical components of the ETSI NFV architecture. With regard to our proposal to extend intelligence to the NFVI-PoPs, we will refer to them as I-PoPs. Both I-PoP #1 and I-PoP #2 (see Fig. 3), which host the vCU and vDU, respectively, are two major building blocks of the infrastructure layer of the proposed RAN slicing architecture, as illustrated in Fig. 1 and described in Section II. The management, orchestration, and allocation of virtual compute, storage, and networking resources of the vCU and vDU in I-PoP #1 and I-PoP #2 are accomplished by their respective hypervisors (see Fig. 2). However, the overall management and orchestration of the virtual resources of a RAN slice in I-PoP #1 and I-PoP #2, the interconnection of both I-PoPs via the middlehaul interface, and the interconnection of the virtualized part with the cellular infrastructure via the fronthaul interface are executed by the VIM. The physical resources of the RU are managed by the 3GPP-NSMS, more precisely by the NG-RAN NSSMF. To allocate, manage, and orchestrate the physical and virtual resources of the radio functionalities of $n \in \mathbb{N}^+$ eMBB, URLLC, and mMTC types of RAN slices in the underlying infrastructure layer, the unified framework of the NFV-MANO and 3GPP-NSMS illustrated in Fig. 2 shall

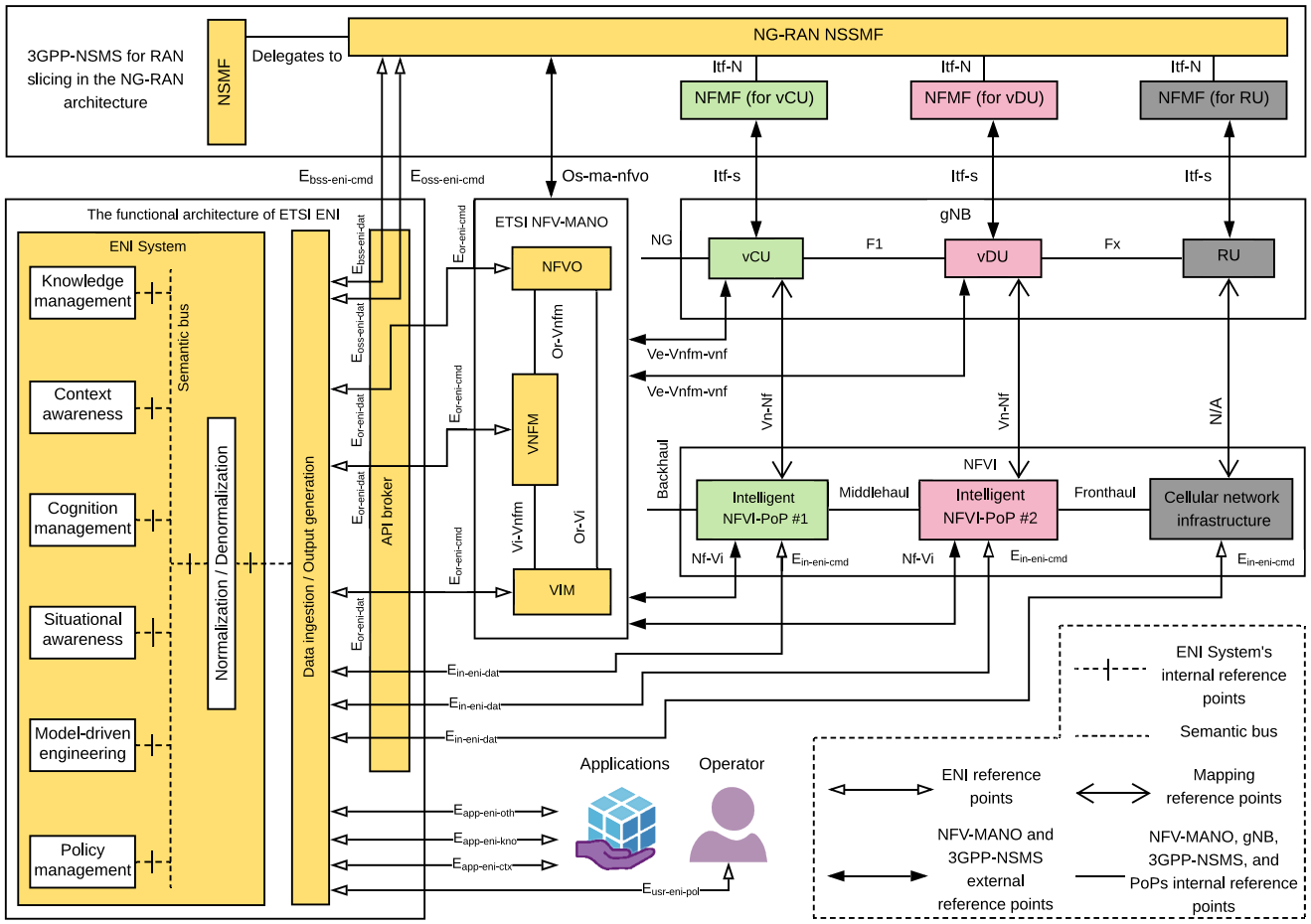

FIGURE 3. **The integration of the ENI System into the unified functioning architecture of the NFV-MANO, 3GPP-NSMS, and I-PoPs proposed for the automation of the management and orchestration, mapping, resource allocation, etc. of RAN slices.**

dynamically harmonize their interactions aiming to provision an efficient RAN slicing.

The vCU, vDU, and VLs of a RAN slice demand dedicated virtual compute, storage, and networking resources in I-PoPs. These virtual resources must be allocated in complete isolation from the virtual resources of other RAN slices operating over shared I-PoPs and transport networks. In order to enhance the efficiency of the allocation, management, and orchestration processes of the proposed architectural framework, the NFV-MANO uses machine-processable descriptors related to the virtual compute, storage, and networking resources of the vCU and vDU throughout the lifetime of a RAN slice. These descriptors, which are the main constituents of the VNFD of vCU and VNFD of vDU, are used among the FBs of the NFV-MANO and during the interaction with the 3GPP-NSMS entities and I-PoPs. We discuss the description models related to such virtual resources of a RAN slice in greater details in Section IV.

D. THE INTEGRATION OF THE ENI SYSTEM INTO THE UNIFIED ARCHITECTURAL FRAMEWORK OF NFV-MANO, 3GPP-NSMS, AND I-POPS

The integration of the ENI System's functioning architecture into the unified framework of the NFV-MANO, 3GPP-NSMS, and I-PoPs (see Fig. 2) is proposed in Fig. 3. The ENI System is integrated via external reference points into the 3GPP-NSMS entities, NFV-MANO FBs, I-PoP #1, I-PoP #2, and cellular network sites. Furthermore, it is also connected to a set of applications that are required by the ENI System to participate in the automation of various operations associated with RAN slices. The primary goals of the integration of the ENI System into these assisted systems are to introduce intelligence and provide automation to the management and orchestration of various types of RAN slices in the NG-RAN architecture. The ENI System must also be controlled by an administrator located within the operator's trusted domain via an external reference point, as shown in Fig. 3. The functioning architecture of the ENI System is composed of an API Broker, Data Ingestion and Output Generation FBs, Normalization and Denormalization FBs, six internal entities, and internal and external reference points [26].

The Data Ingestion FB is an input-facing FB of the ENI System architecture. It is used to ingest structured, semi-structured, and unstructured data – provided via streaming, batch, or on-demand mechanisms – coming from 3GPP-NSMS entities (related to the management and orchestration of the physical components of a RAN slice),

NFV-MANO FBs (related to the management and orchestration of the virtual components of a RAN slice), I-PoPs (related to the underlying virtual resources of a RAN slice), and cellular network infrastructure (related to the underlying radio resources of a RAN slice). The Data Ingestion FB performs filtering, correlation, cleansing, anonymization, pseudonymization, augmentation, and labeling operations on raw data collected from the assisted systems. Once the data collection and ingestion processes from the aforementioned three assisted systems have been completed, the ingested data is then sent to the Normalization FB for interpretation and normalization into a single common and unified data format that is understandable for further processing by the internal entities of the ENI System. On the one hand, the normalization of ingested data related to the RAN slice into a standard format enables network operators to quickly learn about the optimal parameters of each of the nodes in assisted systems; on the other hand, it reduces complex computational problems.

Normalized data coming from the Normalization FB is then routed towards the Semantic Bus, where it is filtered by the Knowledge Management entity. Along with the Knowledge Management entity, the ENI System is composed of five additional internal entities that are based on the observe-orient-decide-act process. They are: Context Awareness, Cognition Management, Situational Awareness, Model-driven Engineering, and Policy Management. These six internal entities (shown in Fig. 3) generate recommendations, predictions, commands, and knowledge for the ENI System in order to assist the management and orchestration related operations of NFV-MANO, 3GPP-NSMS, and the underlying infrastructure for RAN slices in the NG-RAN architecture. These internal entities make use of pre-existing knowledge and/or acquire new knowledge in order to enable the ENI System to adapt its behavior in response to dynamic changes in assisted systems. When the commands and recommendations are generated by the internal entities, they are delivered by the Knowledge Management entity to the Denormalization, which is an output-facing FB of the ENI System.

The Denormalization FB performs the inverse operations of the Normalization FB. It specifically means that when Denormalization FB receives the processed data from the Knowledge Management entity, it is used to convert the processed data coming from the internal entities of the ENI System into a format(s) that is understandable by the assisted systems. The converted data is then delivered to the NFV-MANO, 3GPP-NSMS, and I-PoPs by the Output Generation FB. The assisted systems leverage the commands, predictions, and recommendations of the ENI System to automate the management and orchestration, among other tasks, of the RAN slices in the NG-RAN architecture. The Denormalization FB communicates with the assisted systems through an API broker in order to translate the APIs of the ENI System to the APIs of the assisted systems (see Fig. 3). The API Broker must adhere to the internal and external reference points of the ENI System. The API Broker also defines the proper way for one assisted system (for example, the NFV-MANO) to request services (or assistance) from another assisted system (for example, the 3GPP-NSMS) without requiring the ENI System to understand the details of the APIs of the assisted systems that communicate with each other.

Despite the novel implications of the ENI System in terms of automation and intelligence on the operations of the 3GPP-NSMS, NFV-MANO, and I-PoPs in the NG-RAN architecture, there are a number of research issues that remain unresolved and must be addressed in future reports of the ENI ISG. First and foremost, the internal reference points between the internal entities and the FBs of the ENI Systems are not yet clearly defined. There is an urgent need for such a study in order to design a unified functioning architecture for the ENI System that is compatible with a broad spectrum of assisted systems and applicable to a wide range of use-cases, including various aspects of of RAN slicing in the NG-RAN architecture. Furthermore, the integration of the ENI System into the assisted systems – specifically the 3GPP-NSMS, NFV-MANO, and I-PoPs – is still in its early stages. It will necessitate extensive research efforts to determine the integration of the ENI System into each of the FBs (entities) of the aforementioned assisted systems, investigate the operations (or services) that are needed to be automated, study the implications of the automation of the operations on the performance of the assisted systems, and define their relevant internal and external reference points.

IV. DEFINING AND MODELING THE MAPPING PROBLEM OF THE vCU, vDU, AND VLs OF A RAN SLICE ONTO I-POPS AND TRANSPORT NETWORKS IN THE NG-RAN ARCHITECTURE

To this end, we have provided a detailed insight into the virtual components of a RAN slice. We have also proposed an ENI-assisted architectural framework, which is capable of autonomously hosting, managing, and orchestrating the virtual components of a RAN slice onto the underlying infrastructure in the NG-RAN architecture. Following these discussions, in this section, we model an ENI-enabled functioning framework in order to define the mapping problem of a RAN slice at an architectural level. To accomplish this task, we will first present the mapping process of the vCU and vDU. Next, we will address the mapping process of the internal and external VLs. Then, we will discuss the VNFFGs for the eMBB, URLLC, and mMTC types of RAN slices. We will also address the internal domains of an I-PoP and provide an in-depth discussion of their virtual and physical components. Finally, we will present several assumptions, the main objectives, and a number of constraints that are critical to take into consideration prior to proposing an architectural solution to the mapping problem of a RAN slice.

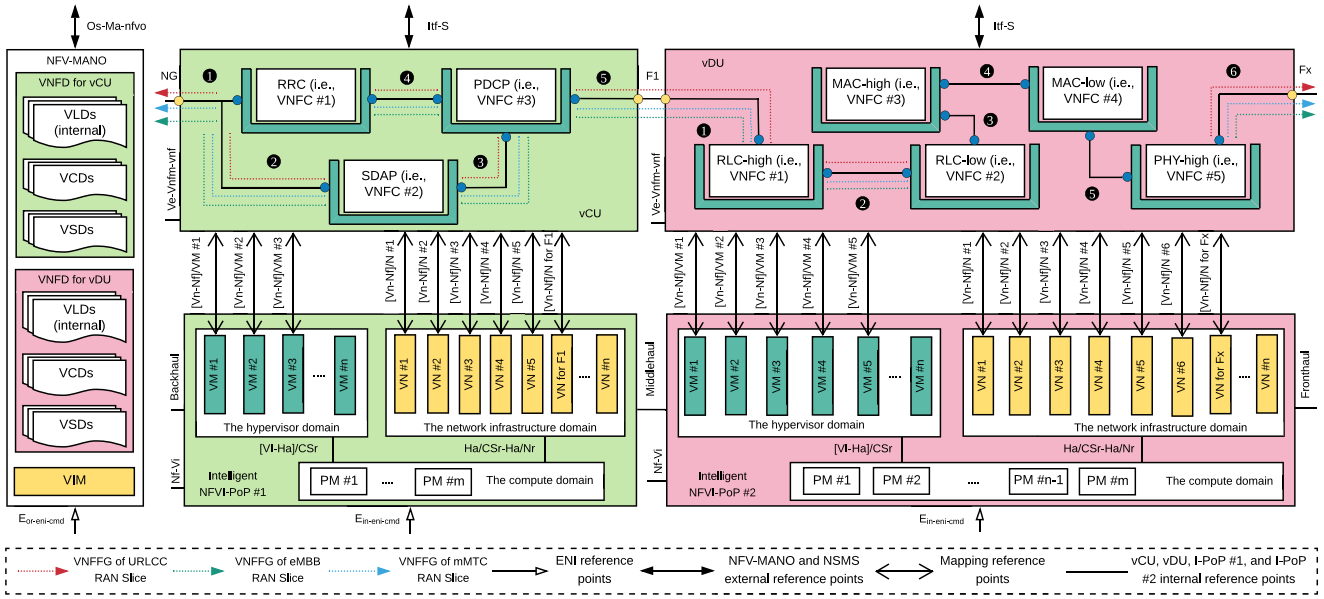

FIGURE 4. **The proposed architecture for automating the management and orchestration of RAN slices (see Fig. 3) is reduced to a reference point architecture that only shows the mapping of the VNFCs of vCU and vDU, internal VLs, and external VLs in both I-PoPs. The VNFCs and VLs are mapped onto VMs and VNs using a one-to-one mapping approach via standard interfaces. These standard interfaces, connecting the various domains of the NFV-MANO with the vCU and vDU, are illustrated in the figure. Due to space constraints, VL #3, VL #4, and VL #5 of vDU are intentionally left blank in the VNFFGs.**

A. DEFINING THE MAPPING PROBLEM OF THE vCU AND vDU

In Section I, we assumed that the vCU and vDU are the VNFs of a RAN slice. We also discovered that a vCU and a vDU have three and five internal functionalities, respectively (see Fig. 1). These internal functionalities are referred to as virtual network function components (VNFCs) in this article. According to the ETSI ISG NFV, the VNFC is a subset of a VNF that performs a well-defined application within the scope of that VNF [89]. Based on this, the vCU has three VNFCs (i.e., VNFC #1 – VNFC #3) that deploy the RRC, SDAP, and PDCP functionalities, respectively, as illustrated on the upper side of Fig. 4. Similarly, the vDU is composed of five VNFCs (i.e., VNFC #1 – VNFC #5), each of which implements the RLC-high, RLC-low, MAC-high, MAC-low, and PHY-high functionalities, respectively. This internal partitioning of vCU and vDU leads to effectively managing the components of a RAN slice, optimally allocating virtual resources to such components, and efficiently enabling and disabling of RAN slice-specific functionalities in a gNB [90]. Additionally, it enables the NFV-MANO to deploy, scale, and upgrade the VNFCs of vCU and vDU independently. Rather than scaling the complete VNF in a RAN slice, only the relevant VNFC within that VNF is scaled.

Furthermore, each VNFC is assumed to be instantiated on a single VM. It is also worth nothing that a VNF might theoretically contain several critical VNFCs that require strict isolation and tight security [91], [92]. To instantiate them, the ETSI proposed the hardware-mediated execution enclave (HMEE) in [93]. Each HMEE is used to host such a critical VNFC in order to protect it from other VNFCs of a VNF in the form of hardware isolation. To ensure compliance with Net Neutrality regulations, we assume that all VNFCs are treated equally in terms of their security and isolation [94]. To that end, we avoid detailed discussion of HMEE in this article. Hence, we assume that all VNFCs within the vCU and vDU, without loss of generality, are instantiated on VMs using a one-to-one mapping approach in this article. This is also illustrated on the upper side of Fig. 4, where the VNFC #1 – VNFC #3 of vCU are mapped onto VM #1 – VM #3 in I-PoP #1, respectively. Similarly, the VNFC #1 – VNFC #5 of vDU are mapped onto VM #1 – VM #5 in I-PoP #2, respectively.

The left side of Fig. 4 illustrates the VNFD of a vCU as well as the VNFD of a vDU. The VNFD defines the deployment and behavior requirements of the vCU and vDU. It contains information about the network connectivity, interfaces, resources, and performance requirements of a VNF. The NFV-MANO FBs use the VNFD to make sure that the VNFCs of the vCU and vDU are placed on the right VMs in the I-PoPs. Each VNFD, among other description information, is made up of a virtual computing descriptor (VCD) and a virtual storage descriptor (VSD). The VCD contains information related to the required amount of virtual compute resources of the vCU and vDU. The VSD references the amount of virtual storage resources that are required by the vCU and vDU.

To specify the amount of virtual compute and storage resources at the VNFC level, the ETSI ISG NFV proposed the virtualization deployment unit (VDU) in [24]. The VDU acts as a descriptor file for a VNFC in the proposed

architecture. It is used to describe the amount of virtual compute and storage resources, which are required by a VM to host its associated VNFC. The VDU is a critical component of the VNFD. By properly defining the parameters and specifying their values of the VDUs of a VNF, the VNFCs of the vCU and vDU can be placed onto suitable VMs, enabling a more reliable, efficient, and performant mapping process of a RAN slice onto the underlying I-PoPs in the NG-RAN architecture.

B. DEFINING THE MAPPING PROBLEM OF THE EXTERNAL VLs AND INTERNAL VLs

Each VNFC has internal connection points that are used to establish internal virtual connections between other VNFCs within the same VNF [12]. When two connection points of two VNFCs of a given VNF are connected, an internal VL is established. This is also highlighted in dark blue color on the upper side of Fig. 4, where the three VNFCs of the vCU and the five VNFCs of the vDU are connected using five and six internal VLs, respectively. Additionally, the vCU and vDU have external virtual connection points (highlighted in yellow color) that are used to establish external virtual connections between the vCU and 5GC, the vCU and vDU, and the vDU and RU via external VLs, namely the next-generation (NG), F1, and Fx interfaces, respectively. The external VL is created by linking two virtual connection points of two VNFCs belonging to two different VNFs [95].

In both intra-and inter-VNF connectivity scenarios, each VL is assumed to be mapped onto a single VN using a one-to-one mapping approach. On the basis of these assumptions, the five internal VLs of vCU and F1 are mapped onto six VNs in I-PoP #1. Similarly, the six internal VLs of vDU and Fx are mapped onto seven VNs in I-PoP #2 (see Fig. 4). The deployment of the internal VLs at each I-PoP, as well as the deployment of the external VLs between I-PoP #1, I-PoP #2, and the cellular network site, should be coordinated in such a way as to deliver an E2E unified RAN slice. The mapping of the NG interface is beyond the scope of this article.

It is worth noting that the NFVO maintains information about the logical connectivity (reachability) within an I-PoP, between the I-PoPs, and the overall networking topology of the underlying transport infrastructure. This information enables the NFVO to determine the appropriate I-PoPs (as well as physical paths on middlehaul and fronthaul) for mapping the internal and external VLs while taking into account the underlying infrastructure's placement constraints. This information is also used by the NFVO to help with the networking resource orchestration of the RAN slice. The NFVO provides network connectivity information to the VIM to deploy and manage the virtual networking resources of the internal VLs in an I-PoP. However, the information related to the external VLs is provided to the Wide area network Infrastructure Manager (WIM) in order to deploy and manage the virtual networking resources over the backhaul and fronthaul links in the NG-RAN architecture. The WIM is a specialized type of VIM that is standardized by the ETSI to manage the virtual networking resources of the external VLs between the I-PoPs and between the I-PoPs and the cellular network sites. For the sake of simplicity, we assume only the VIM for the management and orchestration of the virtual networking resources of both internal and external VLs of a RAN slice in this article.

The VNFDs, depicted on the left side of Fig. 4, are also composed of VLDs that gather information related to the virtual networking resources of the vCU and vDU, as well as their interconnection with the RU in a gNB. The information contained in a VLD is utilized by the NFVO to determine the optimal mapping of a VL instance. In addition, it is also used by the VIM to determine and manage the required virtual networking resources associated with a VL instance in an I-PoP. Each VLD defines the basic topology of the network connectivity within or between the VNFs of a RAN slice, as well as additional parameters (e.g., bandwidth, latency, quality of service (QoS) class, connectivity type, security, etc.) with their values that describe the required performance of a VL. These performance parameters are included in all VLDs. Their values, however, vary depending on the type of RAN slice. There are two types of VLDs: the external VLD and the internal VLD. The former is used to describe the topology and virtual networking resources that are required to connect the 5GC and vCU, the vCU and vDU, and the vDU and RU of a RAN slice. The latter is used to describe the topology and virtual networking resources that are required to connect the VNFCs whiten the vCU and vDU of a RAN slice. It should be noted that the external VLDs are the components of the NSD (as shown on the left side of Fig. 2), whereas the internal VLDs are components of the VNFD (see the left side of Fig. 4).

C. DEFINING THE EMBB, URLLC, AND MMTC TYPES VNFFGS

Once the VNFCs and VLs (internal and external) of a RAN slice have been mapped onto VMs and VNs, the next step is to determine the routes over which the requested traffic should flow – both in downstream and upstream directions. To that end, the VNFFG was proposed by ETSI ISG NFV in [24]. The VNFFG is an effective tool to design, deploy, and manage a RAN slice. To map the VNFFG of a RAN slice successfully, the placement of its VNFCs and VLs onto their respective virtual components should meet the performance requirements while also minimizing the total cost of deployment. It is also used to define the type and amount of requested traffic – matching certain criteria – intended to flow from a source through VNFCs and VLs to a destination in a RAN slice [96]. The traffic flow is controlled by a forwarding path, which is an element of the VNFFG. Each VNFFG may contain at least one forwarding path, which represents the path taken by actual traffic flow on a VL.

Such behavioral and deployment information of a VNFFG is defined in a customized template called the VNFFGD [97].

Similar to other descriptors, the VNFFGD is created in a data modeling language and is subsequently on-boarded into the VNFFG catalog [97]. The VNFFG of a RAN slice is rendered into SFC and a classifier. The former (as standardized by the IETF) creates an ordered list of VNFCs through which traffic should flow from a source to a destination in a forwarding path [55], [98]. The latter is used to classify various types of traffic in accordance with tenant preferences and to filter traffic that should flow through the VNFCs and VLs [99].

There are three types of VNFFGs used to efficiently manage traffic flow in a gNB, as shown in Fig. 4: the eMBB type VNFFG, the URLLC type VNFFG, and the mMTC type VNFFG. Each type of VNFFG must satisfy the internal and external connections, traffic type, and packet flow requirements of its respective RAN slice. It must also clearly illustrate the end-to-end topology, priority dependency between the VNFCs, and overall forwarding rules of the vCU, vDU, and their communication channels with the 5GC and RU in order to meet the demands of the requested traffic [100]. Lastly, the VNFFG must also specify which VNFC (or a feature thereof) of vCU and vDU should be enabled or disabled in order to efficiently meet the performance requirements of a RAN slice. This feature also leads to indicating the exact amount of virtual resources required by a RAN slice in the eMBB, URLLC, and mMTC use case scenarios.

Once the aforementioned information has been specified in the VNFFGDs (of the eMBB, URLLC, and mMTC types of RAN slices) and the VNFFGDs have been uploaded to the network service catalog, the VIM must always use these descriptors to command I-PoP #1 and I-PoP #2 in order to configure and instantiate the VMs and VNs of the respective RAN slice instances [101]. It is worth noting that a RAN slice typically employs a single point-to-point VNFFG, while in some complex use case scenarios, multiple VNFFGs may be deployed to instantiate a RAN slice [102]. In either of the scenarios, if the VNFFG is known ahead of time, the mapping of a RAN slice is a static optimization problem. Conversely, if the VNFFG is unknown, the mapping of the RAN slice becomes a dynamic optimization problem.

D. THE INTERNAL DOMAINS AND COMPONENTS OF AN I-POP

We have previously stated that an I-PoP is a geographical location where the vCU and vDU, as well as the VLs, of a RAN slice are hosted by the VMs and VNs, respectively. The required virtual compute and storage resources of the VNFs, as well as the required virtual networking resources of the internal VLs, are abstracted from the underlying physical compute node and networking links of an I-PoP, respectively [103]. The external VLs of a VNF, on the other hand, are abstracted from the underlying transport links (namely the backhaul, middlehaul, and fronthaul). These transport links connect the I-PoPs and are considered the components of the infrastructure layer [104], as shown in Fig. 1. Should the need arise, the transport links must migrate data and resources of the VMs and VNs of a RAN slice from one I-PoP to another without recreation, re-entering data descriptions, and significant modification of the application(s) being transported [105].

Each I-PoP is divided into three domains: the compute domain, the hypervisor domain, and the network infrastructure domain. The goal of such a partition of an I-PoP is to effectively control and administer the complexity of the infrastructure layer [103]. Although there is always some functional overlap between these domains, they are largely distinct at functional, structural, and practical levels. The positioning of these domains within our proposed architectural framework is shown in Fig. 4 and is discussed further below.

1) THE COMPUTE DOMAIN

Is composed of a set of industry-standard high-volume physical machines (PMs), network interface cards (NICs), input/output accelerators, and storage building blocks [106].

The PMs, along with storage and peripheral equipment, are housed in a server chassis in the form of rack/blade-servers [107]. Each PM, also referred to as a compute node or physical server, is composed of high-speed single/multi-core(s) CPUs and random-access memory (RAM) [107]. These physical resources are first virtualized into virtual central processing units (vCPUs) and virtual random-access memorys (vRAMs), and then allocated to VMs (based on a dynamic resource allocation algorithm) in order to execute the codes of the VNFCs [106]. Each VM is thus equipped with a certain number of vCPUs and vRAMs. The virtualization of a PM into a number of VMs is accomplished in conjunction with a hypervisor that is located within the hypervisor domain [103]. To that end, the three VNFCs of vCU and the five VNFCs of vDU require three and five VMs, which are abstracted from underlying PMs and are subsequently allocated through hypervisor domains in I-PoP #1 and I-PoP #2, respectively, as shown in Fig. 4.

The NIC and input/output accelerators are used to connect the PMs, provide network input/output functionality to the CPUs of PMs, and connect other equipment within the compute domain [108]. The NIC is also virtualized into a set of virtual network interface cards (vNICs) using a hypervisor. Each vNIC is used by a VM to represent its configuration when communicating with other VMs [109]. A VM can be configured to have a single or multiple vNIC(s), each of which is connected to a different VM. However, the traffic between the VMs is connected by a vSwitch that is abstracted from the underlying physical switch [109]. Furthermore, the NIC is responsible for providing physical connectivity between the compute domain and network domain of an I-PoP. In this context, the compute domain uses this type of interconnection exclusively for its internal communication. It does not necessarily support E2E network connectivity between the components of a RAN slice in the NG-RAN architecture [103].

The storage node, which may be coupled with a PM or deployed separately in a storage chassis, stores permanent and temporary data of the VNFCs in the form of hard disk drives, solid state disks, or hybrid disk drives [103]. Each storage node is distinguished by its capacity and a specific level of latency associated with accessing a state held in storage in order to execute an instruction cycle, as well as security, resiliency, cost, and the volatility of the storage [103], [110]. To meet the virtual storage resource requirements of the VMs, the storage nodes are also virtualized using a hypervisor into virtual memory (vMemory) resources. The vMemory could be defined during the creation phase of a VM. It could also be changed or reconfigured during the operation phase in order to accommodate the dynamic changes in virtual resource requirements of a VNFC.

2) THE HYPERVISOR DOMAIN

Is a software environment that is used to abstract virtual compute and storage resources from the underlying physical compute domain, mediate these virtual resources to the VMs, and provide management interfaces to the VIM for the management, orchestration, load-balancing, and monitoring of the VMs [111]. In addition, the hypervisor domain provides an emulated vNIC(s) to each VM allocated to the VNFCs of the vCU and vDU. The vNICs are connected to the vSwitches, located within the hypervisor domain, in order to provide connectivity among the VMs as well as between the VMs and the physical NIC located in the PM.

The VM, vNIC, and vSwitch provided by the hypervisor domain must be equivalent to their original physical environments in terms of performance, functionality, and resource efficiency. Furthermore, the hypervisor domain provides a mechanism for VM migration (running on the same or different PMs or I-PoPs) in order to ensure the atomicity of VM instances [112]. To effectively execute the aforementioned tasks (among others), the hypervisor domain provides the VIM with a list of predefined performance metrics related to the above operations at regular intervals in order to produce real-time and accurate predictions. This significantly improves the performance of VMs, increases their resource utilization ratio, and reduces their energy consumption.

3) THE NETWORK INFRASTRUCTURE DOMAIN

Contains a large number of high-volume switches that are connected to a reliable transport network in order to provide underlying networking resources for the internal and external VLs of a RAN slice [113]. More specifically, it establishes virtual communication channels between the 5GC and vCU, the VNFCs of the vCU and vDU, the vCU and vDU, and the vDU and RU. The virtual communication channels are abstracted from the underlying physical networking resources by utilizing the resource routing and sharing control layer of the virtualization layer that exists within the domain and are subsequently provided in the form of VNs to their respective VLs. Each VL is obtained by establishing a virtual connection between the vNICs of at least two VNFCs. In order to logically connect the VNFCs within/between the vCU and vDU, their respective vNICs need to be configured and then connected via a vSwitch. The network infrastructure domain provides the VIM with a detailed view and reports on virtual and physical nodes aimed at managing and administering these operations and network equipment. These reports and requests are analyzed by the VIM in order to manage and orchestrate the virtual resources – among other operations – associated with inter and intra RAN slice(s) in both I-PoP #1 and I-PoP #2 in the NG-RAN architecture.

E. THE ASSUMPTIONS, OBJECTIVES, AND CONSTRAINTS OF THE MAPPING PROCESS OF A RAN SLICE

This subsection makes several assumptions (at the technical and application levels) in order to derive valid conclusions and correct inferences from the mapping process of the virtual components of a RAN slice. Following that, we present the main objectives we are attempting to achieve through the utilization of the unified functioning architecture of the NFV-MANO, 3GPP-NSMS, the underlying infrastructure, and ENI System. Then, we provide a number of constraints that limit the scope of the proposed mapping architectural solution. These aspects of the mapping process must be clearly defined in the service level agreement (SLA) and continuously monitored for agreement violations during the lifetime of the requested RAN slice. If these requirements are not met, both the service provider and the tenant should include an appropriate penalty value in the SLA. When a service provider fails to deliver assured services, the tenant should impose the penalty according to the agreement.

1) ASSUMPTIONS

For the sake of clarity, we will first make a number of reasonable assumptions prior to mapping the virtual components of a RAN slice onto the underlying infrastructure. These assumptions are as follows:

- We assume the mapping of a single RAN slice in this article. To accomplish this, the RAN NSSMF must first ask the NFVO to instantiate a vCU, a vDU, as well as internal and external VLs, all of which are associated with the requested RAN slice. Notably, we make no assumptions regarding the SST of the requested RAN slice. It could be any of the three SSTs discussed in Section II.
- Following the instantiation of the VNFs and VLs of the requested RAN slice, the associated descriptors (such as VNFDs, VNFFGD, NSD, internal and external VLDs, NSD, and so on) are derived from their respective catalogs and delivered to the NFV-MANO FBs. These descriptors are assumed to comprise (among other attributes) the approximate amount of virtual resources required for each of the VNFCs and VLs based on the SST of the requested RAN slice.
- We assume that the VNFFG of the requested RAN slice has been determined in advance. This specifically

TABLE 3. The list of performance objectives (i.e., metrics) that could be considered for optimization during the mapping process of the virtual components of a RAN slice onto I-PoPs and transport networks in the NG-RAN architecture.

Category	Metric	Category	Metric	Category	Metric
Compute and storage resources	Instantiated VMs of vCU and vDU	Energy consumption	Power consumption of a PM	Service level agreement	Priority
	Active PMs in I-PoP #1 and I-PoP #2		Required power of a VM		Preferences
	VMs resource utilization rate		Power consumption of a VM		Security
	PMs resource utilization rate		VM/PM/I-PoP power wastage		Secracy
	Virtual/physical compute resource usage		Carboon footprint		Privacy
	Virtual/physical storage resource wastage		Thermal dissipation		Network scalibility
	VNFC resource demands prediction		Energy consumption of a VNF and VL		Elasticity
	Intra RAN slices resource allocation	Business	CAPEX		Availability
	VM, PM, and I-PoP stress level		OPEX		Runtime of mapping algorithm
	PM shareability/anti-shareability		Revenue		API authorization level
	Reserved VMs and PMs		Cost		Area of services
	VMs colocation/anti-colocation		Profit		SLA violation
	Virtual and physical resources load balance		Price		Multi-tenancy
Networking resources	VLs resources demand prediction	Migration	VM migration time	RAN slice specific	Downlink per UE/RAN slice
	Jitter, latency, packet loss, and delay		Migration cost		Uplink per UE/RAN slice
	Reserved VLs and PLs		Faiulre rate		Isolation level
	VL and PLs utilization rate		Response time		Maximum number of UEs
	VLs and PLs resource wastage and usage		Number of migration of VMs		Radio spectrum range
	Backhaul/midlehaul/fronthaul distance		Migration overhead		(De)Modulation scheme
	Path redundancy		Inter-VM dependency		(De)Coding scheme
	VL and PL reliability		Interference among VMs		Mobility management
	Transmission cost		Geographical location		Position accuracy
	Path length		Migration acceptance ratio		Mobility interruption time

means that the VNFCs of the vCU and vDU, as well as the internal and external VLs connecting the VNFCs, are anticipated to be processed sequentially and chained exactly as illustrated in Fig. 4.

- Each VNFC requires a fixed number of virtual compute and storage resources. Each VL, whether it is internal or external, utilizes a certain amount of bandwidth. The amount of virtual resources required by these components is assumed to vary over the lifetime of the requested RAN slice. Hence, we assume that it is a dynamic optimization problem from a resource allocation perspective.
- The VNFCs of a VNF can be hosted by the same or different PMs. In the latter case, the bandwidth of each internal VL is limited to a fixed amount, whereas in the former case, each internal VL has an infinite bandwidth [114]. Additionally, the operations management, maintenance, and resource orchestration are not sophisticated in the first scenario. On the basis of these advantages, we assume that the VNFCs of vCU and vDU of the requested RAN slice are placed on the same PM in I-PoP #1 and I-PoP #2, respectively.

2) OBJECTIVES

Based on the aforementioned assumptions, the primary goal of the mapping of the requested RAN slice onto I-PoPs and transport networks is to find an optimal solution aimed at optimizing (maximizing, minimizing, or balancing) a set of predefined performance objectives (i.e., metrics), which are listed in Tab. 3. These objectives, which are grouped into seven categories, are used to evaluate the performance of a successful mapping solution of a RAN slice using publicly available analytical and simulation tools. The mapping of a RAN slice can be a single-objective or multiple-objective optimization problem. In the former case, a single objective can be optimized during a specified time interval; in the latter case, multiple objectives can be optimized during a specified time interval. The following are some common objectives:

- One of the objectives is to keep the number of active PMs in an I-PoP to a minimum. This can be accomplished efficiently by utilizing state-of-the-art ML-assisted algorithms to predict the virtual resources required for the VNFCs. The underlying PMs can then be partitioned into multiple VMs, each of which is configured and allocated autonomously to its respective VNFC based on resource predictions. This results in avoiding under-utilization and over-utilization of virtual resources and minimizing resource wastage in an I-PoP.
- The second common objective is to keep the total bandwidth consumption of the underlying PLs in transport networks to a minimum. To accomplish this objective, the required virtual networking resources for each VL in the requested RAN slice can be predicted in the first stage using ML-assisted algorithms. Following that, the PL must be partitioned in such a way that the total

required bandwidth of the VLs does not exceed the total bandwidth of the host PL. Finally, the VL must be mapped autonomously onto the most suitable PL in the transport network.

- Reducing the number of VM and VN migrations from an over-loaded PM and PL to a less-loaded PM and PL is another common objective that can be achieved by the proposed unified functioning architecture. To accomplish this task, a real-time ML-assisted migration algorithm is required that (a) predicts the future resource demands of VMs or VNs based on their historical resource requirements, (b) places VMs and VNs on nodes and links with available resources, and (c) monitors the resource consumption of the nodes and links on a regular basis. By automating these tasks, the number of migrations of virtual components of a RAN slice is expected to decrease to a minimum.
- Decreasing the power consumption of I-PoPs and transport networks is one of the critical performance objectives of the mapping process of a RAN slice. Among other factors, the energy consumption of I-PoPs and transport networks can be increased primarily due to the high volume of communication traffic, the number of active PMs and PLs, inefficient utilization of resources, and cooling equipment. Thus, autonomous management of the components of the underlying infrastructure through the use of ML-assisted algorithms in the context of the proposed solution can significantly reduce energy consumption.

3) CONSTRAINTS

The mapping problem of the VNFCs and VLs of vCU and vDU can be limited by a number of intrinsic and use-case-specific constraints that must be stated prior to proposing the architectural solution. Some of the most critical constraints that we believe have a significant impact on the quality of the performance objectives of the mapping process are enumerated below. It should be noted that the performance objectives (i.e., metrics) listed in Tab. 3 can also be used as constraints when formulating the mapping problem.

- The virtual compute and storage resource demands of a VNFC must not exceed the virtual compute and storage resource capacity of the host VM. Likewise, the sum of the virtual compute and storage resource demands of all VMs belonging to the VNFCs of a VNF (either vCU or vDU) of the requested RAN slice must not exceed the physical compute and storage resource capacity of the host PM in an I-PoP.
- The virtual networking resource demands of a VL should not exceed the virtual networking resource capacity of the host VN. Similarly, the sum of the virtual networking resource demands of the internal and external VLs belonging to a VNF (either vCU or vDU) of the requested RAN slice should not exceed the total available physical bandwidth of the underlying host PL in the transport network.
- Each VM and VN is limited to hosting a single type of VNFC and VL at a given time, respectively. For example, the VM hosting the MAC-high should be configured exclusively to host this particular VNFC throughout the lifetime of the requested RAN slice. This type of customization of VM and VN results in two advantages: (a) If performance degrades or the RAN slice fails, the root causes of the failure can be identified and resolved quickly and effectively. (b) The VNFCs and VLs can be shared between multiple RAN slices with identical or dissimilar SSTs, thereby optimizing the trade-off between isolation level, resource utilization, and customization.
- To avoid service interruptions and improve survivability, redundant VMs and VNs may be configured in each of the I-PoPs and transport network. The redundant VMs and VNs must be instantiated in the event that any of the activated (or primary) VMs or VNs fails, with the goal of rapidly recovering the requested RAN slice [115]. Due to these advantages, the redundant VM should not be configured on the same PM as the primary VM is hosted. Likewise, the redundant VN must not be configured on the same physical path as the primary VN is hosted.

V. THE PROPOSED ARCHITECTURAL SOLUTION FOR MAPPING THE vCU, vDU, AND VLs OF A RAN SLICE ONTO I-POPS AND TRANSPORT NETWORKS IN THE NG-RAN ARCHITECTURE

Following the definition and modeling of the mapping problem in the previous section, we propose an architectural solution in this section that aims to jointly map the VNFs and VLs of a single RAN slice onto I-PoPs and transport networks in an automated fashion in the NG-RAN architecture. The solution leverages the ENI System in order to optimize the performance of the mapping process by utilizing cutting-edge ML-assisted techniques. The proposed solution executes the mapping process in three distinct phases: (a) the resource automation phase, which applies intelligence and automation to the mapping process of a RAN slice; (b) the resource management phase, which manages and orchestrates virtual resources for inter-and intra-RAN slices; and (c) the resource allocation phase, which allocates VMs to the VNFCs and VNs to the VLs of a RAN slice. These three phases are carried out by three sets of major building blocks, which are connected via standard reference points. The broad design of the proposed solution is depicted in Fig. 5. Its three phases, as well as their customized architectures derived from Fig. 5, are addressed in detail in the following subsections, respectively.

A. THE VIRTUAL RESOURCE AUTOMATION PHASE

The functioning architecture of the virtual resource automation phase is depicted in Fig. 6, which is derived from the general architecture of the proposed mapping solution in Fig. 5. This phase is executed by the ENI System. To begin,

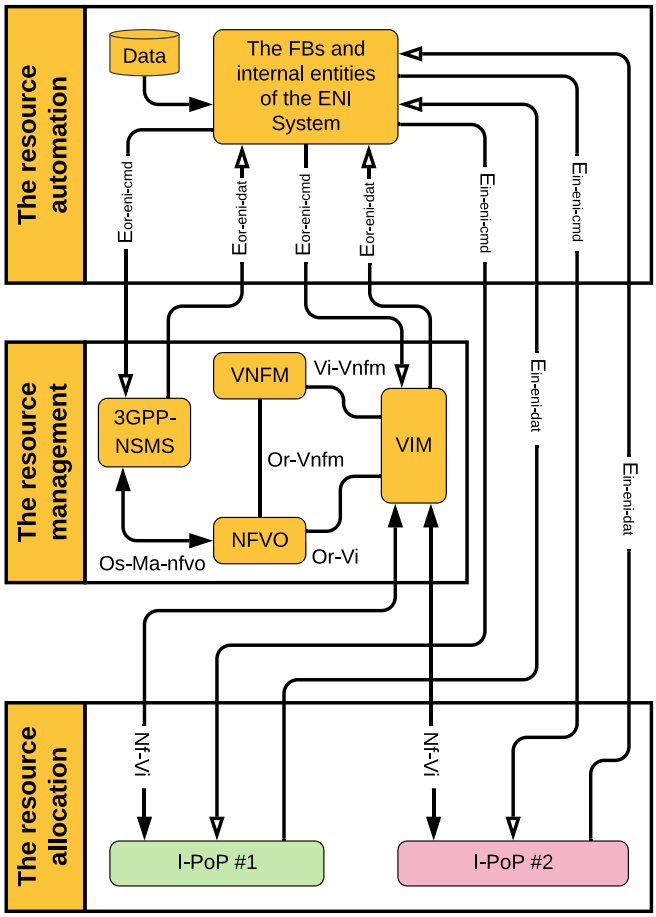

FIGURE 5. The three major sets of building blocks of the proposed architectural framework for mapping the vCU, vDU, and VLs of a RAN slice onto I-PoPs and transport networks in the NG-RAN architecture. The legends contained in this figure have been defined in the previous sections.

the proposed architectural framework collects appropriate historical data – at a faster rate, with a higher sense of accuracy, and on a larger scale – pertaining to three categories: (a) the requirements of the eMBB, URLLC, and mMTC types of RAN slices; (b) the design and engineering of the three types of RAN slices; and (c) the calculation of the allocated, available, and reserved virtual resources in I-PoPs and transport networks (see step 1 in the resource automation phase of Fig. 6). The ENI System collects these three types of data from the 3GPP-NSMS, NFV-MANO, and the underling infrastructure. From these three categories of historical data, the proposed framework selects only the data that is relevant to a set of performance objectives we are attempting to optimize during the mapping process of the requested RAN slice. These performance objectives include minimizing the number of active PMs, reducing the number of active PLs, decreasing the number of VM and VN migrations, and lowering energy consumption, among others.

The historical data (or collected data) is then cleaned, randomized, and visualized using a variety of data cleaning, randomization, and visualization techniques. It then partitions the collected data into two distinct sets: the training set and the evaluation set (steps 2-3). The training set is used to train the mapping model data (offline) in order to generate accurate predictions for a number of performance metrics, such as the required amount of virtual resources, the level of isolation, the end-user's (tenant's) preferences, and so on [82]. The proposed framework selects an appropriate ML-assisted algorithm prior to training the mapping model, which must analyze the historical data successfully before proceeding with the mapping procedure. The evaluation data set is utilized to assess the performance objectives of the chosen ML-assisted algorithm with respect to the mapping process throughout the lifetime of the requested RAN slice [82].

At this stage of the mapping process, it is assumed that the ML-assisted model has been successfully trained (offline) and is ready to be deployed (online) to the unified architectural framework of the NFV-MANO, 3GPP-NSMS, I-PoPs, and ENI System (step 4). First, the offline trained mapping model is transferred to the Cognition Management FB, which uses it to generate predictions (or recommended actions) about the requirements and design of the requested RAN slice, as well as the status of virtual resources in underlying I-PoPs (step 5). The offline trained mapping model is then transferred to the Knowledge Management FB, where it is kept for periodic (or continuous) learning (step 6). This periodic learning is accomplished through the utilization of real-time data acquired from the assisted systems and is intended for the autonomous incremental development of the respective ML-assisted mapping model [26], [82].

It is worth stating once more that the proposed architectural framework acquires real-time data for a variety of performance metrics associated with the requested RAN slice from 3GPP-NSMS (specifically, the NG-RAN NSSMF), NFV-MANO (specifically, the VIM), and the underlying infrastructure (specifically, the I-PoP #1 and I-PoP #2). The NG-RAN NSSMF provides the ENI System with information about the physical components and their connections to the virtual components of the RAN slice (step 7). The VIM provides information about the design of VNFFG, the required quantity of virtual (compute, storage, and networking) resources, and the performance requirements of the RAN slice (step 7). The I-PoPs provide information on the status of the virtual resources that are provided to the VNFCs and VLs of the vCU and vDU of the requested RAN slice, such as the number of (available, allocated, and consolidated) PMs (or VMs) and PLs (or VLs), network throughput, and so on (step 7).

The three types of assisted systems described above may provide data to the ENI System's Data Ingestion FB in a variety of formats, with different ranges of features and varying levels of structures. As a result, the Data Ingestion FB first programmatically filters and normalizes this input (raw) data into a uniform (or standardized) data format before passing it on to the other FBs of the ENI System for further processing (step 7). Choosing the appropriate data normalization technique (such as min-max, z-score, mean, and so on) is critical

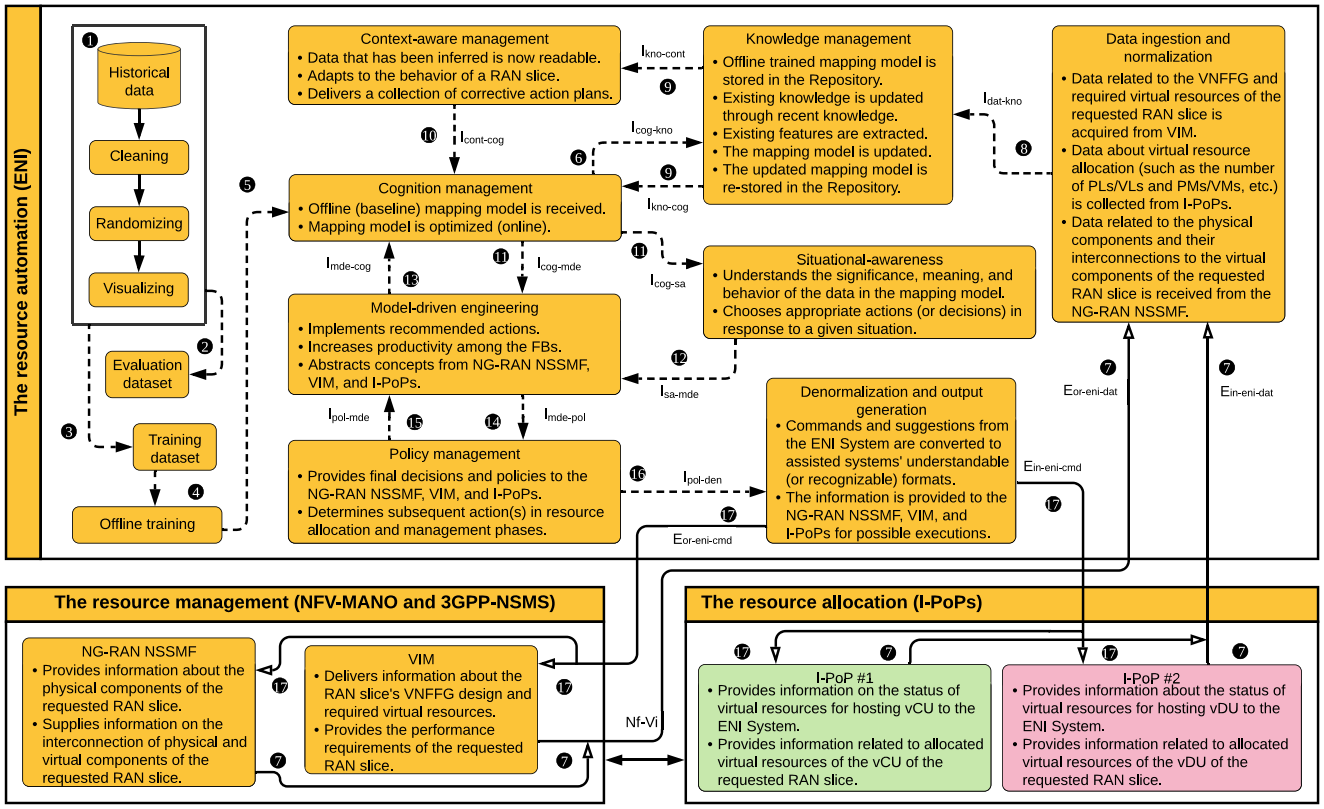

FIGURE 6. **The proposed functioning architecture for the virtual resource automation phase of the mapping process of vCU, vDU, and VLs of a RAN slice onto I-PoPs and transport networks in the NG-RAN architecture. It should be noted that the resource automation phase is executed by the ENI System, and its architectural framework is derived from the functioning architecture of the proposed mapping solution illustrated in Fig. 5. The legends contained in this figure have been defined in the previous sections.**

for improving the model's accuracy in mapping the virtual components of the requested RAN slice. The normalized data (a type of data in which the values of numeric columns in a dataset are accurately converted to a common scale) is subsequently transferred to the Knowledge Management FB (step 8).

The ML-assisted mapping model, which has already been developed through offline training on a pre-prepared dataset and is stored in the Knowledge Management FB (see step 6), is updated on a regular (or as needed) basis with newly arrived input data from the Normalization FB. These input data may result in certain changes (modifications) to the existing mapping model, such as feature extraction, editing the current knowledge, adding new knowledge, and so on. Once these modifications have been made, an updated version of the mapping model is created and stored in the Knowledge Repository of the Knowledge Management FB. This process of optimizing (or re-training) the existing mapping model is performed continuously throughout the lifetime of the requested RAN slice, which is tuned by the real-time stream of data coming from the NG-RAN NSSMF, VIM, and I-PoPs. The overarching goal of this optimization is to increase the accuracy of the selected mapping model. The proposed framework then transfers the optimized mapping model (also known as inferred data) to the Context-aware Management FB and Cognition Management FB concurrently (step 9).

The Context-aware Management FB is the first to read the newly arrived inferred data from the Knowledge Management FB (step 9). Furthermore, the Context-aware Management FB must ensure that the inferred data is stored in standardized formats that must be readable by the rest of the FBs of the ENI System. The inferred data, which pertains to the state and environment of the performance objectives (see Tab. 3) of the mapping process, may change throughout the lifetime of the requested RAN slice. As a result, the Context-aware Management FB is used to continuously monitor (or measure) the performance objectives and adapt the behavior of a RAN slice in accordance with the SLA [8]. If we assume that an unusual event is detected (for example, a VM allocated to a VNFC is overloaded or a VN assigned to an internal-VL is underutilized), this FB, in conjunction with other FBs of the ENI System, may take corrective actions in accordance with a set of pre-established rules in order to resolve such a hazardous situation. Fig. 6 illustrates that the Context-aware Management FB passes context information related to the requested RAN slice to the Cognition Management FB (step 10).

The Cognition Management FB enables the ENI System to analyze normalized ingested data, context, and information associated with the mapping model of the requested RAN

slice. Once such an understanding of the mapping model is accomplished, this FB assesses the acquired data and defines the actions that must be taken in order to meet the performance objectives of the mapping process (step 9). The Cognition Management FB is also in charge of making predictions about the QoS and quality of experience (QoE) metrics associated with mapping the virtual components of the RAN slice. These predictions are achieved through the use of offline data (gathered during offline training), inferences (provided by online training), and context information. The probable values generated by the mapping algorithm for each of the predicted parameters enable the network operator to forecast the likelihood of the mapping process of the requested RAN slice. Once the desired predictions have been generated, the Cognition Management FB delivers them to the Situational-awareness FB and Model-driven Engineering FB at the same time (step 11).

Through the use of Situational-awareness FB, the proposed architectural framework is enabled to fully comprehend and analyze the significance, meaning, and behavior of all events that have occurred or will occur during the mapping process of the requested RAN slice in the NG-RAN NSSMF, VIM, and I-PoPs. This FB also enables the proposed framework to understand how the requested RAN slice's behavior influences the performance objectives that the ENI System is attempting to optimize in the short, medium, and long term. Furthermore, in collaboration with the Model-driven Engineering FB, the Situational-awareness FB determines what the ENI System should do in response to the given event(s), makes (or chooses) the most appropriate decisions, and performs the relevant action (or combination of actions). These decisions or recommended set of optimal actions are then forwarded to the Model-driven Engineering FB (step 12).

The main objective of the Model-driven Engineering FB in proposed framework is to make appropriate decisions aimed at implementing recent recommendations or sets of actions produced by the Situational-awareness FB. To achieve that goal, this FB utilizes model-driven engineering mechanisms in order to convert these actions into a form of policy, such as employing machine-readable models rather than code. In addition, the Model-driven Engineering FB employs a collection of algorithms to abstract information and concepts in order to manage the behavior of the requested RAN slice intelligently in the NG-RAN NSSMF, VIM, and I-PoPs. This FB also boosts productivity across the FBs of the ENI System by reusing the standardized models. For example, it assists the Cognition Management FB in making predictions and the Knowledge Management FB in measuring the performance objectives of the requested RAN slice (step 13). Fig. 6 shows that the converted form of policies is then passed to the Policy Management FB (step 14).

The Policy Management FB provides scalable and consistent decisions that must be made during the resource management and resource allocation phases in order to guarantee the key performance indicators (KPIs) of the performance metrics of the requested RAN slice, such as PM consolidation, VM allocation, VM migration, virtual resource reservation for emergency cases, and so on. The Policy Management FB also decides on the next action(s) based on the performance of the previous action(s) taken by the NG-RAN NSSMF, I-PoPs, and VIM. If the performance of the action is not satisfactory, this FB adjusts its configuration, redefines (or edits) the relevant policies, and then proceeds with future action(s) under specified conditions (step 15). Fig. 6 illustrates that the Policy Management FB then forwards the policy and final decisions of the mapping process to the Denormalization and Output Generation FB (step 16).

Once the predictions, commands, recommendations, and/or suggestions related to the resource management and resource allocation phases are generated by the internal FBs of the ENI System, the final step (during the resource automation phase) is to provide them to the NG-RAN NSSMF, VIM, I-PoP #1, and I-PoP #2 aimed at intelligently optimizing the performance objectives of the mapping process of the requested RAN slice. To accomplish this, such information is first converted by the Denormalization and Output Generation FB into standardized format(s) that must be understandable by the aforementioned assisted systems. Following that, as shown in Fig. 6, the ENI System's recommendations, commands, and predictions are provided to the assisted systems in order to intelligently execute and implement the management, orchestration, and allocation of virtual resources of the virtual components of the RAN slice (step 17). On the basis of these commands, the NG-RAN NSSMF, VIM, and I-PoPs are anticipated to take the necessary course of actions aiming to automate the tasks related to the mapping process of the vCU, vDU, internal VLs, and external VLs of the requested RAN slice onto the underlying physical infrastructure in the NG-RAN architecture.

B. THE VIRTUAL RESOURCE MANAGEMENT PHASE

Fig. 7 depicts the proposed functioning architecture for the phase of virtual resource management and orchestration. It is derived from the general architectural framework of the mapping process, which is shown in Fig. 5. As illustrated in Fig. 7, the NG-RAN NSSMF and NFV-MANO FBs are in charge of the virtual resource management and orchestration phase of the proposed mapping solution. The NG-RAN NSSMF, which serves as an input-facing entity to the NFV-MANO FBs, makes use of the RAN NSST to define the required resources for a RAN slice. The NFV-MANO FBs, which are the main building blocks of this phase, rely on a variety of NFV description files in order to instantiate, allocate, and manage virtual resources throughout the lifetime of the requested RAN slice [116].

At the start of this phase, the NG-RAN NSSMF checks the SST and SD of the requested RAN slice (step 1 in the virtual resource management phase of Fig. 7). We assume that the tenant requests only a single RAN slice. Therefore, the SD is by default set to one. Once the SST has been determined

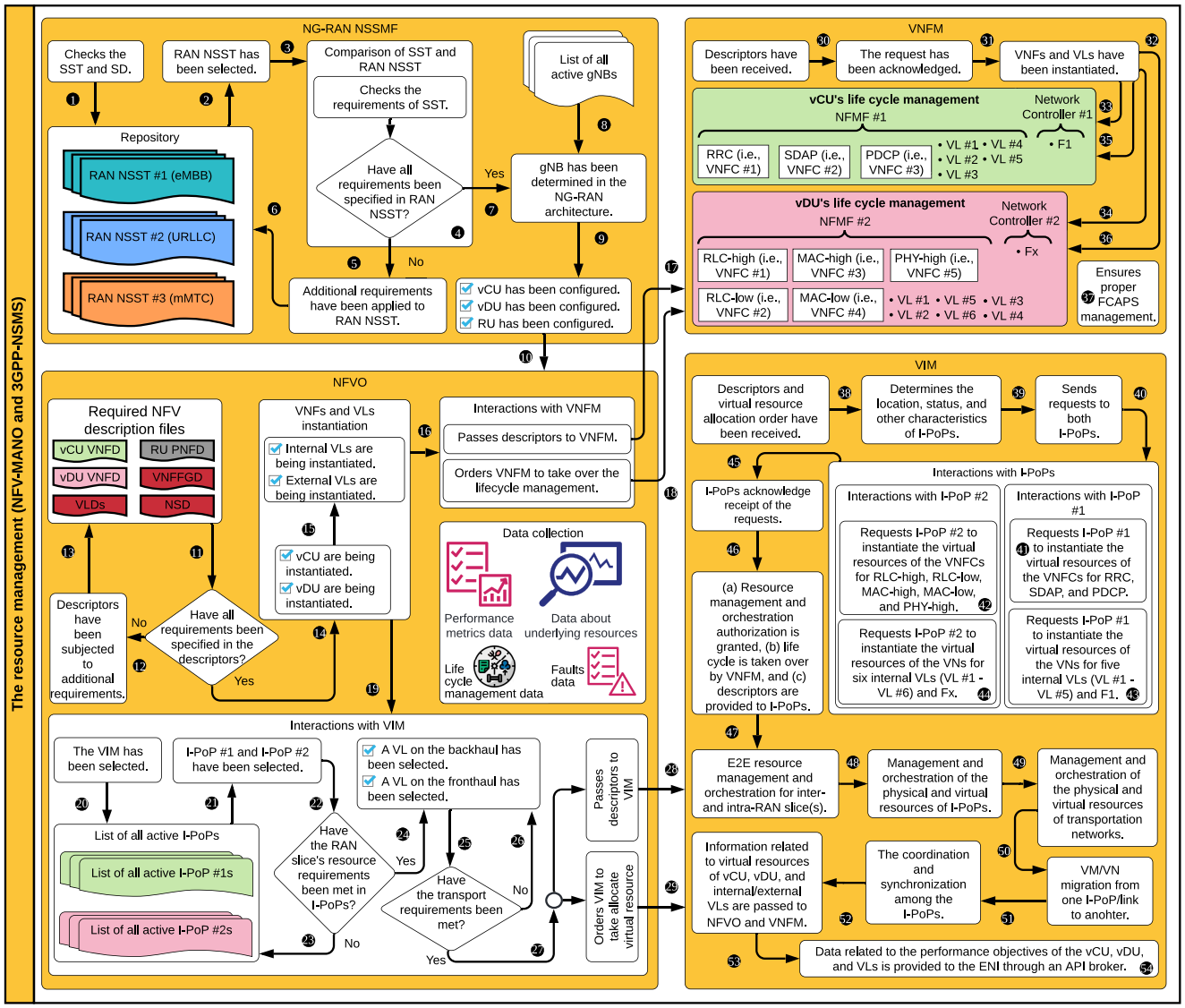

FIGURE 7. The proposed functioning architecture of the virtual resource management and orchestration phase. This framework is derived from the mapping process's overall architectural solution (see Fig. 5). The NG-RAN NSSMF and NFV-MANO FBs execute this phase of the mapping framework. The legends contained in this figure have been defined in the previous sections.

(step 2), the NG-RAN NSSMF selects the RAN NSST of the requested RAN slice from the repository (step 3). We make no assumptions regarding any particular SST. Hence, the proposed solution is applicable to the mapping process of any of the 3GPP-defined SSTs. The RAN NSST contains all optional and mandatory information related to the configuration parameters, functionalities, and features of the requested RAN slice [117].

The NG-RAN NSSMF then compares the requirements specified in the requested SST to those defined in the RAN NSST to ensure that an appropriate RAN NSST has been selected from the repository (step 4). If we assume that some requirements are not properly defined or the tenant requests the inclusion of additional optional requirements (such as full or partial access to billing and/or charging information, priority, service exposure capabilities, and many others), the RAN NSSMF must re-define the selected RAN NSST (step 5) and upload it to the repository for future use (step 6). However, if all requirements of the requested SST have been defined in the selected RAN NSST in a satisfactory manner (step 7), the RAN NSSMF proceeds to determining the gNB that must cover the desired geographical area of the RAN slice (step 8). The NG-RAN NSSMF selects this gNB from a list of active gNBs in the NG-RAN architecture. It also configures and adjusts the behavior of the vCU, vDU, and RU in accordance with the requirements of the RAN slice specified in the RAN NSST (step 9) [117]. Finally, the RAN NSSMF provides information related to the required virtual resources (among other parameters) to NFVO in order to enable the deployment of the vCU and vDU, as well as to manage and orchestrate their virtual resources throughout the lifetime, of the requested RAN slice (step 10).

Following that, the NFVO prepares and on-boards the required descriptors of vCU and vDU based on the information received from the RAN NSSMF (step 11). They primarily consist of VNFDs for the vCU and vDU, internal and external VLDs, PNFD for the RU, NSD, and VNFFGD. The NFVO shall compare the elements of these descriptors to the requirements associated with the virtual components of the requested RAN slice as provided by the RAN NSSMF. In the event that the descriptors do not contain all of the requirements (step 12), the NFVO re-defines the elements or adds new ones and subsequently uploads modified NFV descriptors to their respective repositories (step 13). Nevertheless, if the descriptors contain all the requirements and their elements are appropriately defined, the NFVO proceeds with the instantiation of the virtual components of the requested RAN slice (step 14). During the instantiation process, we let the NFVO begins the instantiation of the vCU and vDU (step 15), followed by the instantiation of their internal and external VLs (step 16). To that end, the NFVO transfers the description files to the VNFM (step 17) and instructs the VNFM to take over the tasks associated with the instantiation and life cycle management for the vCU, vDU, and internal and external VLs of the requested RAN slice (step 18).

The NFVO then determines the VIM that must provide access to both I-PoPs and transport networks in order to manage and orchestrate the underlying virtual compute, storage, and networking resources of inter-RAN and intra-RAN slices (step 19). Once the VIM has been determined (step 20), the NFVO selects the I-PoPs that will accommodate the vCU, vDU, and their associated internal and external VLs of the requested RAN slice. The NFVO selects two appropriate I-PoPs from a list of active I-PoPs, taking into account a variety of technical, regulatory, and geographic constraints (step 21). Next, the NFVO compares the vCU and vDU resource requirements to the available, reserved, and allocated virtual resources in both I-PoPs (step 22). If the I-PoPs do not meet the resource requirements of the requested RAN slice, the NFVO searches for other appropriate I-PoPs that meet those requirements (step 23). In the event that the selected I-PoPs are able to meet the resource requirements, the NFVO proceeds with the orchestration of connectivity between the I-PoPs as well as their connectivity to the cellular network sites across the transport networks (step 24). Hence, the NFVO selects appropriate VLs on the backhaul and fronthaul in order to connect vCU to vDU and vDU to RU, respectively (step 25). If the transport network requirements of the requested RAN slice have not been met by the selected VLs (step 26), the NFVO searches for and selects other appropriate VLs across the backhaul and fronthaul networks. However, if these requirements are met (step 27), the NFVO prepares to interact with the VIM for the allocation of the required virtual compute, storage, and networking resources. Finally, the NFVO transfers the description files to the VIM (step 28) and orders the VIM to initiate the required virtual resource allocation for the vCU, vDU, and VLs (step 29).

It is also worth noting that the NFVO collects data on performance metrics, resources, faults, and life cycle management from the VIM and VNFM on a periodic basis in order to dynamically scale (both horizontally and vertically) vCU, vDU, and internal and external VLs throughout the lifetime of the requested RAN slice.

Next, we assume that the VNFM received the required description files and an order from the NFVO (step 30) to configure, instantiate, and take over life cycle management of the virtual components of the requested RAN slice. Although the instantiation of VNFs and internal VLs is initially within the scope of VNFM, this procedure begins at the NFVO with the selection of appropriate description files and the execution of several prerequisite tasks, such as assigning VNFs and internal VLs to an associated VNFM, specifying their lifetime, defining specific features and functionalities for their life cycle management, and many others (see steps 15 and 16). It is worth noting that the instantiation and life cycle management of the external VLs are beyond the scope of the VNFM. The WIM and Network Controller are responsible for these two tasks related to the external VLs of a RAN slice over the midlehaul and fronthaul transport links [118].

After the VNFM and WIM acknowledge the request (step 31), the NFVO grants the VNFM and WIM access to data repositories in order to proceed with the instantiation of the VNFs and VLs, as well as manage their life cycles. The VNFM and WIM make use of the description files to complete the remaining tasks associated with the instantiation of VNFs and VLs. First, we let VNFM complete the instantiation process of the VNFCs of vCU (RRC, SDAP, and PDCP), the VNFCs of vDU (RLC-high, RLC-low, MAC-high, MAC-low, and PHY-high), as well as the internal VLs of vCU and vDU. Then, we let WIM complete the instantiation process of the external VLs (i.e., F1 and Fx). We assume that the VNFM and WIM successfully instantiated the vCU and vDU, as well as their internal and external VLs, of the requested RAN slice in I-PoPs and transport links (step 32).

Regarding life cycle management, the VNFM assigns a dedicated NFMF #1 to the vCU and its internal VLs (step 33), as well as a dedicated NFMF #2 to the vDU and its internal VLs (step 34). Likewise, the WIM assigns Network Controller #1 to F1 (step 35) and Network Controller #2 to Fx (step 36). The VNFM assigns a unique identification number (ID) to the life cycle management operations of the requested RAN slice. The NFV-MANO FBs, as well as the I-PoPs and transport links, use this ID to identify all operations and resources pertaining to the requested RAN slice. The VNFM and WIM must also guarantee that the FCAPS of vCU, vDU, and VLs are properly managed and the NFVO is notified on a regular basis (step 37). Due to the scope of the paper, we skip providing further details on workflow messages exchanged between the NFVO, VNFM, and WIM.

Following the instantiation of the VNFs and VLs, we assume that the VIM selected by the NFVO is fully aware of

the SST, description files, life cycle management, and virtual resource requirements of the requested RAN slice (step 38). The VIM also knows the geographical locations, identities, resource status, and connectivity information of both I-PoPs that host the vCU and vDU (step 39). Based on this knowledge, the VIM sends virtual resource allocation requests (in coordination with the NFVO and VNFM) to both I-PoPs (step 40). These requests also include resource management and allocation constraints applicable to the required resources of the vCU and vDU. The VIM utilizes these constraints to determine available resource zones within an I-PoP, as well as to perform optimal partition, reservation, and allocation of the underlying resources. To ensure synchronization between the vCU and vDU of the requested RAN slice, the VIM must send these requests to both I-PoPs at the same time. First, the VIM requests I-PoP #1 and I-PoP #2 to configure, instantiate, and allocate the virtual compute and storage resources for the three VNFCs of vCU (step 41), as well as for the five VNFCs of vDU (step 42). The VIM then requests I-PoP #1 to configure, instantiate, and allocate the virtual networking resources for five internal and one external VLs of vCU (step 43). It also requests I-PoP #2 to configure, instantiate, and allocate the virtual networking resources for six internal and one external VLs of vDU (step 44).

Once the I-PoPs receive the virtual resource allocation requests (step 45) and acknowledge receipt of these requests (step 46), the VIM (a) requests permission from the NFVO to manage and orchestrate the required virtual resources; (b) notifies the VNFM to take over the life cycle management of these resources; and (c) transfers all necessary information related to IDs, lifetime, connectivity, and description files to both I-PoPs in order to proceed with the configuration and allocation of the required resources (step 47).

Despite the fact that the I-PoPs are responsible for configuring and allocating virtual resources, a number of higher-layer tasks are still in the scope of VIM. These tasks, which are performed sequentially after resource instantiation, include the following: (a) E2E management and orchestration of the virtual resources of the requested RAN slice (step 48); (b) management and orchestration of the underlying virtual and physical resources of the I-PoPs (step 49), as well as the transport networks (step 50); (c) migration (if necessary) of the VMs and VNs from I-PoP #1 to I-PoP #2 and vice versa (step 51); and (d) coordination and synchronization among the I-PoPs (step 52) [77]. Furthermore, the VIM collects and periodically passes information regarding the configuration, allocation, and reservation of virtual resources of the vCU, vDU, and internal and external VLs to the NFVO and VNFM (step 53). This information is used by both of these FBs to dynamically optimize the operations associated with the life cycle management and virtual resources of the RAN slice.

The VIM additionally gathers and passes data related to the performance metrics (see Tab. 3) of the vCU, vDU, and internal and external VLs to the ENI System through an API broker (step 54). Once the input data from the VIM has been processed and the recommendations (or commands) have been generated by the internal FBs of the ENI System (see the preceding subsection), the ENI System passes such recommendations back to the VIM with the goal of automating and intelligently performing the tasks within the scope of the VIM. The VIM uses these recommendations to automate and intelligently manage, among others, the following four major tasks: coordination among the I-PoPs; management and orchestration of the overall operations of the I-PoPs; management and orchestration of virtual and physical resources of the requested RAN slice; and the migration of the VMs and VL from one I-PoP to another.

C. THE VIRTUAL RESOURCE ALLOCATION PHASE

The proposed architectural framework for the virtual resource allocation phase is shown in Fig. 8. Similar to the previous two phases, the functioning architecture of this phase is also derived from the general architectural framework of the mapping process, as depicted in Fig. 5. The I-PoPs (#1 and #2) and transport networks are in charge of executing the phase of virtual resource allocation. They configure, allocate, and terminate virtual compute and storage resources for the VNFCs of the vCU and vDU, as well as the virtual networking resources of the VLs, of the requested RAN slice.

To begin, the I-PoP #1 and I-PoP #2 receive requests from the VIM to prepare, configure, and allocate virtual resources for vCU, vDU, and VLs (step 1 in the virtual resource allocation phase of Fig. 8). After processing the requests and granting permission to host the RAN slice (step 2), each I-PoP notifies its hypervisor regarding the mapping of the VNF and its associated VLs. Based on the predictions generated by the ENI System and according to the description files received from the NFVO, the hypervisors analyze the virtual compute and storage resource demands (workloads) of the VNFCs and the virtual networking resource demands (bandwidth) of the VLs in both I-PoPs and transport networks (step 3). We assume that all PMs and PLs are placed sequentially in I-PoPs and transport networks. The hypervisors then begin the process of configuring virtual compute and storage resources for the VNFCs (step 4).

To that end, the hypervisor (in I-PoP #1 and I-PoP #2) searches for an active resource-sufficient PM whose available physical compute and storage resources must equal or exceed the sum of the virtual compute and storage resource demands of the VNFCs of a VNF (step 5). The hypervisors perform searching for an active PM in a list of sequentially-ordered PMs in I-PoP #1 and I-PoP #2 (step 6). We assume that each hypervisor has selected a PM (step 7). In I-PoP #1, the hypervisor configures the virtual compute and storage resources for three VMs (each for a VNFC of vCU) on the selected PM (step 8). In parallel, the hypervisor configures the virtual compute and storage resources of five VMs (each for a VNFC of vDU) on the selected PM in I-PoP #2 (step 8). It is critical to ensure that the virtual compute and storage resource requirements of a VNFC do not

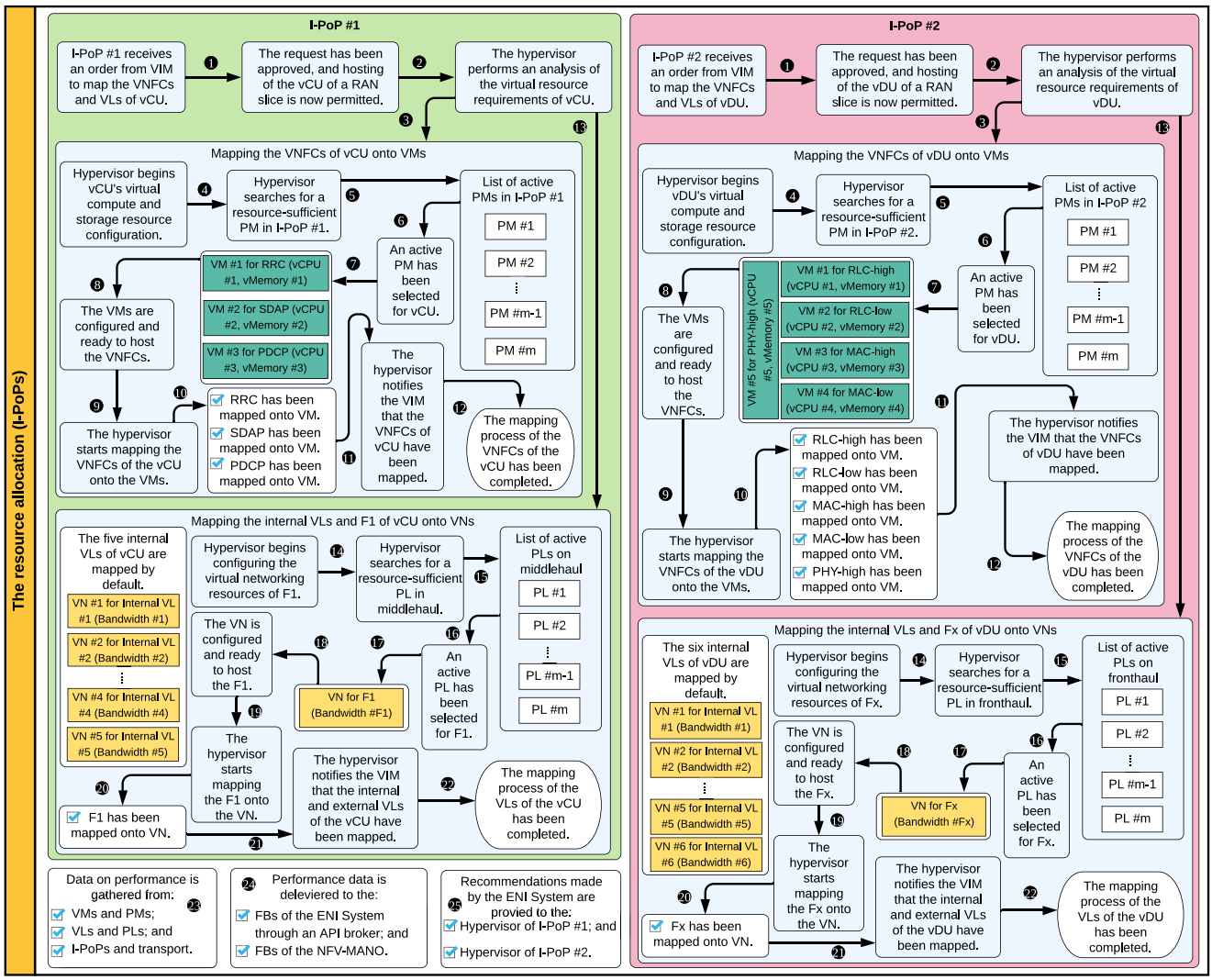

FIGURE 8. **The virtual resource allocation phase's proposed functioning architecture. It is derived from the mapping process's general architectural framework, which is illustrated in Fig. 5. The I-PoP #1 and I-PoP #2 are in charge of the virtual resource allocation phase. The legends contained in this figure have been defined in the previous sections.**

exceed the virtual compute and storage resource capacity of the host VM when configuring the selected PM. We assume that the VMs have been configured and are prepared to host the VNFCs of the vCU and vDU (step 9). The hypervisors thus programmatically begin mapping the VNFCs onto the underlying configured VMs in I-PoP # 1 and I-PoP #2 at the same time (step 10). To accomplish this, it starts with the mapping of the first VNFC and concludes with the mapping of the final VNFC of a VNF (step 11). Such a sequential chaining and processing of the VNFCs of the requested RAN slice must be specified in its corresponding VNFFGD. The hypervisor notifies the VIM of mapping the VNFCs of a VNF (step 12). Finally, the VNFCs have been successfully mapped onto their corresponding VMs in an I-PoP.

Following that, the hypervisors begin mapping the internal and external VLs (step 13). Due to the assumption that the VNFCs of a VNF are hosted on the same PM in an I-PoP, no mapping process is required for the internal VLs of the requested RAN slice. To map F1 and Fx, the hypervisors begin configuring virtual networking resources (step 14). First, a list of active PLs must be searched (step 15) for those with sufficient physical networking resources capable of mapping F1 and Fx onto the middlehaul and fronthaul, respectively (step 16). After selecting a resource-sufficient PL (step 17), the hypervisors configure a VN for F1 and Fx on both the middlehaul and fronthaul, respectively (step 18). When configuring the selected PL, it is critical to ensure that the virtual networking resource requirements of a VL do not exceed the virtual networking resource capacity of the host VN in both middlehaul and fronthaul transport links. The VN is configured and ready to host the VL (step 19). The hypervisors start mapping the F1 and Fx onto their respective VNs sequentially (step 20), as specified in the VNFFGD of the requested RAN slice. The F1 and Fx have been mapped onto VNs (step 21). The VIM is then notified of the F1 and Fx mapping (step 22). Finally, the internal VLs and external

VLs of vCU and vDU have been successfully mapped onto I-PoPs, as well as middlehaul and fronthaul transport links, respectively.

The virtual compute and storage resource requirements of the VNFCs and the virtual networking resource requirements of the VLs may change throughout the lifetime of the requested RAN slice. Hence, the resource capacity of the VMs and VNs shall also be reconfigured by the proposed solution in an automated fashion in response to changes in resource demands, in order to meet the virtual resource requirements of the respective virtual components of the RAN slice. In addition, the hypervisors may configure redundant VMs and VNs to avoid service interruption, improve survivability, and increase availability [115]. Providing redundancy, however, may necessitate allocating and reserving extra virtual resources, thereby increasing total costs and complicating resource management. This may require an agreement between the service provider and tenant in the SLA regarding the additional cost and various levels of redundancy [119]. If both parties agree on providing redundancy, the hypervisors must configure redundant VMs and VNs on different PMs and PLs than those that host the primary VLs and VNs. The configuration of redundant VMs for the VNFCs of vCU and vDU is performed from steps 5 to 8 in I-PoP #1 and I-PoP #2, respectively. The configuration of redundant VNs for F1 and Fx is performed from steps 15 to 19 on middlehaul and fronthaul, respectively.

Once the mapping process of a RAN slice and the configuration of redundant VMs and VNs have been completed, the proposed framework updates the knowledge repositories of the hypervisors in I-PoP #1 and I-PoP #2. Following that, the hypervisors collect data related to the performance objectives (see Tab. 3) from VMs and PMs, VLs and PLs, and other components of I-PoPs and transport networks on a regular basis (step 23). The performance data must contain all necessary information about the status of the VMs, VNs, PMs, and PLs. It must also include the number of hosted vCUs, vDUs, VLs, as well as the available, reserved, and allocated resources within I-PoPs and transport networks. Such a collection of performance data assists the I-PoPs and transport networks in managing and orchestrating their resources, as well as deciding whether to accept or reject future RAN slice requests. The hypervisors then pass performance data to the management entities of the I-PoPs.

The I-PoPs construct performance data into two distinct sets. The first data set is transferred to the VIM (step 24). The VIM utilizes this data to manage the operations and resources of the I-PoPs and transport networks. The VIM also shares this information with the NFVO and VNFM in order to effectively manage and orchestrate the life cycle and resources of the requested RAN slice throughout its lifetime. The second data set is routed through an API broker to the ENI System (step 24). It contains information about the metrics associated with the performance objectives of the requested RAN slice. This data set is then processed by the internal entities and FBs of the ENI System in order to generate recommendations regarding the operations and resources of the I-PoPs (step 25).

To accomplish this, the ENI System employs ML-assisted algorithms to (a) predict the virtual resource requirements of the VMs and VNs of the requested RAN slice based on the historical data or learning from an environment in steps 5 and 15; (b) configure the virtual compute and storage resources of VMs, as well as the virtual networking resources of VNs, in steps 8 and 18; and (c) dynamically allocate them to the VNFCs and VLs in steps 11 and 21. In comparison to the state-of-the-art VNF mapping algorithms, the use of ML-assisted algorithms in the preceding steps of the virtual resource allocation phase is critical for obtaining the performance objectives such as optimizing virtual resource allocation performance, minimizing energy consumption, and intelligently scaling the requested RAN slice [69], [70], [120].

The proposed framework also provides ENI's recommendations for physical resource virtualization. Thus, the hypervisors in I-PoPs and transport networks could efficiently abstract the virtual resources of VMs and VNs from the underlying physical resources of the PMs and PLs based on the predictions generated by the relevant ML-assisted algorithm(s) of the ENI System. At this stage of the virtual resource allocation phase (steps 7 to 9 and steps 17 to 19), accurate predictions of the virtual resource requirements of VMs and VNs of the requested RAN slice are critical for the hypervisor to partition the PMs and PLs and efficiently allocate them to the VNFs and VLs of the requested RAN slice. In comparison to the state-of-the-art resource virtualization and abstraction solutions, such intelligent and automated virtualization of physical resources in I-PoPs and transport networks undoubtedly results in achieving the performance objectives set in the previous section such as the reduction of the active number of PMs and PLs, the lowering of VM and VN migration, the maximization of revenue and profit, and the avoidance of over-and under-utilization of physical resources.

VI. CONCLUSION AND FUTURE OUTLOOK

In this article, we extended the architectural framework of network slicing, defined by the NGMN Alliance, towards the NG-RAN architecture in order to provision the eMBB, URLLC, and mMTC types of RAN slices in 5G and beyond mobile communication networks. Based on this framework, we addressed the management and orchestration of the operations and resources of RAN slices by leveraging the unified functioning architecture of the NFV-MANO, 3GPP-NSMS, and the underlying compute and transport network infrastructure. Then, we integrated the ENI System into the NFV-MANO FBs, 3GPP-NSMS entities, and the components of the underlying infrastructure in order to automate their operations and bring intelligence to the edge of 5G and beyond cellular networks. These contributions were primarily aimed at laying the groundwork for an autonomous and

intelligent mapping of the virtual components of a RAN slice. To that end, we defined and modeled the mapping problem of the vCU, vDU, and VLs, taking into account a number of performance objectives that are limited by a set of constraints. Finally, we proposed an ENI-enabled architectural solution that is both autonomous and intelligent for mapping the VNFCs of vCU and vDU onto their respective VMs, as well as the internal and external VLs of the vCU and vDU onto their corresponding VNs, in I-PoPs and transport links, respectively.

Regarding the future work, there are several challenges that need to be addressed. First, we would like to extend our work on virtualizing and partitioning the physical resources of the I-PoPs and transport network on top of the aforementioned ENI-enabled framework in the NG-RAN architecture. Second, we are interested in looking into how virtual resources are allocated for various types of RAN slices using cutting-edge autonomous and intelligent tools with the goal of optimizing resource utilization and enhancing RAN slice performance while also taking into account a set of predefined constraints, such as the trade-off between resource utilization ratio and isolation level, the harmonization of inter-RAN and intra-RAN slice resource allocation algorithms, and the management of inter-RAN and intra-RAN slice priority. Third, the physical and virtual compute, storage, and networking resources of RAN slices should be managed and orchestrated dynamically and opportunistically across the underlying infrastructure. Therefore, we are also interested in applying cutting-edge ML-assisted algorithms to resource management and orchestration in the near future, as well as exploring the impact of automation and intelligence on these two areas.

GLOSSARY

3GPP	Third Generation Partnership Project
3GPP-NSMS	3GPP-network slicing management system
5G	fifth-generation
5GC	5G core network
5GPPP	Fifth Generation Public-Private Partnership
AI	artificial intelligence
AM	acknowledged mode
API	application programming interface
ARQ	automatic repeat request
CAPEX	capital expenditure
CN	core network
CPU	central processing unit
C-RAN	cloud-RAN
CSMF	communication service management function
CU	centralized unit
CVM	container in virtual machine
DAS	direct attached storage
DL	deep learning
DRB	data radio bearer
DU	distributed unit
E2E	end-to-end
eCPRI	evolved common public radio interface
EM	element manager
eMBB	enhanced mobile broadband
ENI	Experiential Networked Intelligence
ETSI	European Telecommunications Standards Institute
FB	functional block
FCAPS	fault, configuration, accounting, performance, and security
FL	federated learning
F-RAN	fog-RAN
gNB	next-generation NodeB
GSMA	Global System for Mobile Communications Association
GST	Generic Network Slice Template
HARQ	hybrid ARQ
H-CRAN	heterogeneous-cloud RAN
HMEE	hardware-mediated execution enclave
HMTC	high-performance machine-type communications
ID	identification number
IETF	Internet Engineering Task Force
IP	internet protocol
I-PoP	Intelligent PoP
ISG	industry specification group
KB	kilobyte
Kbps	kilo bits per second
KPI	key performance indicator
MAC	medium access control
ML	machine learning
mMTC	massive machine-type communications
MVNO	mobile virtual network operator
NAS	network attached storage
NEST	NEtwork Slice Type
NF	network function
NFMF	network function management function
NFV	Network Function Virtualization
NFVI	network function virtualization infrastructure
NFVI-PoP	network function virtualization infrastructure point of presence
NFV-MANO	network function virtualization-management and orchestration
NFVO	network function virtualization orchestrator
NG	next-generation
NGMN	Next Generation Mobile Networks
NGP	Next Generation Protocols
NG-RAN	next-generation radio access network
NIC	network interface card
NIC	virtual network interface card
non-RT RIC	non-real-time RAN Intelligent Controller
NR	new radio
nrt-RIC	near real-time RAN Intelligent Controller
NS	network slice
NSD	network service descriptor
NSI	network slice instance

NSMF	network slice management function
NSMS	network slicing management system
NSS	network slice subnet
NSSAI	network slice selection assistance information
NSSMF	network slice subnet management function
NSST	network slice subnet template
NST	network slice template
OPEX	operational expenditure
O-RAN	Open-RAN
OSS/BSS	operations support system/business support system
packet	data convergence protocol
PDU	protocol data unit
PHY	physical
PL	physical link
PM	physical machine
PNF	physical network function
PNFD	physical network function descriptor
PoP	point of presence
PS	physical server
QFI	QoS flow ID
QoE	quality of experience
QoS	quality of service
RAM	random-access memory
RAN	radio access network
RF	radio frequency
RL	reinforcement learning
RLC	radio link control
RRC	radio resource control
RRM	radio resource management
RU	radio unit
SAN	storage area network
SCF	Small Cell Forum
SD	slice differentiator
SDAP	service data adaptation protocol
SDC	software-defined compute
SDN	software-defined networking
SDO	standards development organization
SDS	software-defined storage
SFC	service function chain
SLA	service level agreement
S-NSSAI	single NSSAI
SST	slice/service type
TIP	Telecom Infra Project
TM	transparent mode
UM	unacknowledged mode
URLLC	ultra-reliable low latency communications
user	equipment
V2X	vehicle-to-everything
VCD	virtual computing descriptor
vCPU	virtual central processing unit
vCU	virtual centralized unit
vDU	virtual distributed unit
VDU	virtualization deployment unit
VIM	virtualized infrastructure manager
VL	virtual link
VLD	virtual link descriptor
VM	virtual machine
vMemory	virtual memory
VN	virtual networking
VNF	virtual network function
VNFC	virtual network function component
VNFD	virtual network function descriptor
VNFFG	virtual network function forwarding graph
VNFFGD	virtual network function forwarding graph descriptor
VNFM	virtual network function manager
vRAM	virtual random-access memory
VSD	virtual storage descriptor
vSwitch	virtual switch
WIM	Wide area network Infrastructure Manager

ACKNOWLEDGMENT

The authors are grateful to the anonymous reviewers for their insightful comments and valuable remarks, which significantly improved the quality of this article, as well as to the Area Editor and Associate Editor for coordinating the peer-review process.

REFERENCES

[1] *Technical Specification Group Services and System Aspects; Telecommunication Mangement; Study on Management and Orchestration of Network Slicing for Next Generation Network (Release 15), V15.1.0*, 3GPP Standard TS 28.801, Jan. 2018.

[2] L. M. Contreras, C. J. Bernardos, A. de la Oliva, X. Costa-Pérez, and R. Guerzoni, "Orchestration of crosshaul slices from federated administrative domains," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, 2016, pp. 220–224.

[3] *Deliverable D3.2; 5G NORMA Network Architecture; Intermediate Report, Version 1.0*, 5G-NORMA, Heidelberg, Germany, Jan. 2017.

[4] *Deliverable D3.2; 5G-ESSENCE Final Report on Network Embedded Cloud; 5G cSD-RAN Controller and Network Slicing, Version 1.0*, 5G-ESSENCE, Athens, Greece, May 2019.

[5] *Deliverable D1.1; 5G-CHARISMA Intelligent, Distributed Low-latency Security C-RAN/RRH Architecture, Version 1.1*, 5G-CHARISMA, Heidelberg, Germany, Jun. 2016.

[6] X. Li *et al.*, "5G-crosshaul network slicing: Enabling multi-tenancy in mobile transport networks," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 128–137, Aug. 2017.

[7] M. A. Habibi, B. Han, F. Z. Yousaf, and H. D. Schotten, "How should network slice instances be provided to multiple use cases of a single vertical industry?" *IEEE Commun. Stand. Mag.*, vol. 4, no. 3, pp. 53–61, Sep. 2020.

[8] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "The structure of service level agreement of slice-based 5G network," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun.*, Bologna, Italy, Sep. 2018, pp. 1–6.

[9] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[10] F. Giannone *et al.*, "Impact of virtualization technologies on virtualized RAN midhaul latency budget: A quantitative experimental evaluation," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 604–607, Apr. 2019.

[11] H. Hirayama, Y. Tsukamoto, S. Nanba, and K. Nishimura, "RAN slicing in multi-CU/DU architecture for 5G services," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, 2019, pp. 1–5.

[12] "Network functions virtualisation (NFV) release 3; evolution and ecosystems; report on network slicing support with ETSI NFV architecture framework, v3.1.1," ETSI, Sophia Antipolis, France, ETSI Rep. GR NFV-EVE 012, Dec. 2017.

[13] *Technical Specification Group Services and System Aspects; Management and Orchestration; Architecture Framework (Release 17), V17.1.0*, 3GPP Standard TS 28.533, Dec. 2021.

[14] M. A. Habibi, B. Han, M. Nasimi, N. P. Kuruvatti, A. Fellan, and H. D. Schotten, "Towards a fully virtualized, cloudified, and slicing-aware RAN for 6G mobile networks," in *6G Mobile Wireless Networks* (Computer Communications and Networks), Y. Wu *et al.*, Eds. Cham, Switzerland: Springer, 2021, pp. 327–358.

[15] *O-RAN WG1, O-RAN Architecture Description, V05.00*, O-RAN-Alliance, Alfter, Germany, Jul. 2021.

[16] U. U. Rahman, K. Bilal, A. Erbad, O. Khalid, and S. U. Khan, "Nutshell—Simulation toolkit for modeling data center networks and cloud computing," *IEEE Access*, vol. 7, pp. 19922–19942, 2019.

[17] L. Ruiz *et al.*, "Genetic algorithm for holistic VNF-mapping and virtual topology design," *IEEE Access*, vol. 8, pp. 55893–55904, 2020.

[18] *Description of Network Slicing Concept*, NGMN-Alliance, Frankfurt am Main, Germany, Jan. 2016.

[19] "A deliverable by the NGMN alliance: NGMN 5G white paper two, v 1.0," NGMN-Alliance, Frankfurt am Main, Germany, Rep., Jul. 2020.

[20] *NGMN Overview on 5G RAN Functional Decomposition, V1.0*, NGMN-Alliance, Frankfurt am Main, Germany, Feb. 2018.

[21] *5G RAN CU—DU Network Architecture, Transport Options and Dimensioning, V1.0*, NGMN-Alliance, Frankfurt am Main, Germany, Apr. 2019.

[22] *Technical Specification Group Services and System Aspects; Service Requirements for the 5G System; Stage 1. (Release 18), V18.5.0*, 3GPP Standard TS 22.261, Dec. 2021.

[23] *Technical Specification Group Radio Access Network; NG-RAN; Architecture Description (Release 16), V16.8.0*, 3GPP Standard TS 38.401, Dec. 2021.

[24] *Network Functions Virtualisation (NFV); Virtual Network Functions Architecture*, ETSI Standard GS NFV-SWA 001, Dec. 2012.

[25] "Network functions virtualisation (NFV); management and orchestration; network service templates specification, v2.1.1," ETSI, Sophia Antipolis, France, ETSI Rep. GS NFV-IFA 014, Oct. 2016.

[26] "Experiential networked intelligence (ENI); system architecture, v1.1.1," ETSI, Sophia Antipolis, France, ETSI document GS ENI 005, Sep. 2019.

[27] "Cloud architecture and deployment scenarios for O-RAN virtualized RAN," O-RAN-Alliance, Alfter, Germany, Rep. O-RAN WG6 CAD-V01.00.00, Oct. 2019.

[28] A. Garcia-Saavedra and X. Costa-Pérez, "O-RAN: Disrupting the virtualized RAN ecosystem," *IEEE Commun. Stand. Mag.*, vol. 5, no. 4, pp. 96–103, Dec. 2021.

[29] *Network Slicing Use Case Requirements*, GSMA, London, U.K., Apr. 2018.

[30] *GSM Association; Official Document NG.116—Generic Network Slice Template, Version 2.0*, GSMA, London, U.K., Oct. 2019.

[31] *Considerations on Network Virtualization and Slicing*, IETF, Fremont, CA, USA, Nov. 2017.

[32] Y. L. Lee, J. Loo, and T. C. Chuah, "A new network slicing framework for multi-tenant heterogeneous cloud radio access networks," in *Proc. Int. Conf. Adv. Electr. Electron. Syst. Eng. (ICAEES)*, 2016, pp. 414–420.

[33] D. H. Kim, S. M. A. Kazmi, A. Ndikumana, A. Manzoor, W. Saad, and C. S. Hong, "Distributed radio slice allocation in wireless network virtualization: Matching theory meets auctions," *IEEE Access*, vol. 8, pp. 73494–73507, 2020.

[34] V. N. Ha and L. B. Le, "End-to-end network slicing in virtualized OFDMA-based cloud radio access networks," *IEEE Access*, vol. 5, pp. 18675–18691, 2017.

[35] S. Vural, N. Wang, P. Bucknell, G. Foster, R. Tafazolli, and J. Muller, "Dynamic preamble subset allocation for RAN slicing in 5G networks," *IEEE Access*, vol. 6, pp. 13015–13032, 2018.

[36] D. Marabissi and R. Fantacci, "Highly flexible RAN slicing approach to manage isolation, priority, efficiency," *IEEE Access*, vol. 7, pp. 97130–97142, 2019.

[37] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, Jan. 2019.

[38] H. Xiang, W. Zhou, M. Daneshmand, and M. Peng, "Network slicing in fog radio access networks: Issues and challenges," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 110–116, Dec. 2017.

[39] B. Ojaghi, F. Adelantado, A. Antonopoulos, and C. Verikoukis, "SlicedRAN: Service-aware network slicing framework for 5G radio access networks," *IEEE Syst. J.*, early access, Mar. 29, 2021, doi: 10.1109/JSYST.2021.3064398.

[40] A. Laghrissi, T. Taleb, M. Bagaa, and H. Flinck, "Towards edge slicing: VNF placement algorithms for a dynamic & realistic edge cloud environment," in *Proc. IEEE Global Commun. Conf.*, 2017, pp. 1–6.

[41] I. Benkacem, T. Taleb, M. Bagaa, and H. Flinck, "Optimal VNFs placement in CDN slicing over multi-cloud environment," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 616–627, Mar. 2018.

[42] A. De Domenico, Y.-F. Liu, and W. Yu, "Optimal virtual network function deployment for 5G network slicing in a hybrid cloud infrastructure," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7942–7956, Dec. 2020.

[43] H. Gupta, M. Sharma, A. Franklin, and B. R. Tamma, "Apt-RAN: A flexible split-based 5G RAN to minimize energy consumption and handovers," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 1, pp. 473–487, Mar. 2020.

[44] R. L. Gomes, F. R. P. da Ponte, A. Urbano, A. Vasconselos, L. F. Bittencourt, and E. R. M. Madeira, "Energy-aware slicing of network resources based on elastic demand through daytime," in *Proc. IFIP/IEEE Symp. Integr. Netw. Serv. Manag. (IM)*, 2019, pp. 604–608.

[45] S. Ravindran, S. Chaudhuri, J. Bapat, and D. Das, "EESO: Energy efficient system-resource optimization of multi-sub-slice-connected user in 5G RAN," in *Proc. IEEE Int. Conf. Electron. Comput. Commun. Technol. (CONECCT)*, 2020, pp. 1–6.

[46] J. Mei, X. Wang, and K. Zheng, "An intelligent self-sustained RAN slicing framework for diverse service provisioning in 5G-beyond and 6G networks," *Intell. Converg. Netw.*, vol. 1, no. 3, pp. 281–294, Dec. 2020.

[47] T. C. Chuah and Y. L. Lee, "Intelligent RAN slicing for broadband access in the 5G and big data era," *IEEE Commun. Mag.*, vol. 58, no. 8, pp. 69–75, Aug. 2020.

[48] O. Adamuz-Hinojosa, P. Munoz, J. Ordonez-Lucena, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "Harmonizing 3GPP and NFV description models: Providing customized RAN slices in 5G networks," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 64–75, Dec. 2019.

[49] T. Pamuklu, M. Erol-Kantarci, and C. Ersoy, "Reinforcement learning based dynamic function splitting in disaggregated green open RANs," Feb. 2021, *arXiv:2012.03213*.

[50] H. Yu, F. Musumeci, J. Zhang, M. Tornatore, and Y. Ji, "Isolation-aware 5G RAN slice mapping over WDM metro-aggregation networks," *J. Lightw. Technol.*, vol. 38, no. 6, pp. 1125–1137, Mar. 15, 2020.

[51] F. Z. Morais *et al.*, "PlaceRAN: Optimal placement of virtualized network functions in the next-generation radio access networks," 2021, *arXiv:2102.13192*.

[52] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, "Scheduling wireless virtual networks functions," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 2, pp. 240–252, Jun. 2016.

[53] L. Diez, A. M. Alba, W. Kellerer, and R. Agüero, "Flexible functional split and fronthaul delay: A queuing-based model," *IEEE Access*, vol. 9, pp. 151049–151066, 2021.

[54] C. Mei, J. Liu, J. Li, L. Zhang, and M. Shao, "5G network slices embedding with sharable virtual network functions," *J. Commun. Netw.*, vol. 22, no. 5, pp. 415–427, Oct. 2020.

[55] H. Hantouti, N. Benamar, and T. Taleb, "Service function chaining in 5G & beyond networks: Challenges and open research issues," *IEEE Netw.*, vol. 34, no. 4, pp. 320–327, Jul./Aug. 2020.

[56] H. A. Alameddine, S. Sebbah, and C. Assi, "On the interplay between network function mapping and scheduling in VNF-based networks: A column generation approach," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 4, pp. 860–874, Dec. 2017.

[57] J. Martín-Pérez and C. J. Bernardos, "Multi-domain VNF mapping algorithms," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2018, pp. 1–6.

[58] M. A. T. Nejad, S. Parsaeefard, M. A. Maddah-Ali, T. Mahmoodi, and B. H. Khalaj, "vSPACE: VNF simultaneous placement, admission control and embedding," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 542–557, Mar. 2018.

[59] M. Gamal, S. Jafarizadeh, M. Abolhasan, J. Lipman, and W. Ni, "Mapping and scheduling for non-uniform arrival of virtual network function (VNF) requests," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, 2019, pp. 1–6.

[60] Z. Shaoping, G. Xiujiao, and Y. Hongfang, "Virtual network function instantiation and service function chaining mapping in wide area network," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2016, pp. 1–6.

[61] M. Jalalitabar, G. Luo, C. Kong, and X. Cao, "Service function graph design and mapping for NFV with priority dependence," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2016, pp. 1–5.

[62] M. A. Habibi, B. Han, and H. D. Schotten, "Network slicing in 5G mobile communication: Architecture, profit modeling, and challenges," in *Proc. 14th Int. Symp. Wireless Commun. Syst.*, Bologna, Italy, Nov. 2021, pp. 1–6.

[63] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A comprehensive survey of RAN architectures toward 5G mobile communication system," *IEEE Access*, vol. 7, pp. 70371–70421, 2019.

[64] F. Meneses, M. Fernandes, D. Corujo, and R. L. Aguiar, "SliMANO: An expandable framework for the management and orchestration of end-to-end network slices," in *Proc. IEEE 8th Int. Conf. Cloud Netw. (CloudNet)*, 2019, pp. 1–6.

[65] *Technical Specification Group Services and System Aspects; System architecture for the 5G System (5GS); Stage 2 (Release 17), V17.2.0*, 3GPP Standard TS 23.501, Sep. 2021.

[66] W. Shi *et al.*, "Two-level soft RAN slicing for customized services in 5G-and-beyond wireless communications," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4169–4179, Jun. 2022.

[67] J. Zou, S. A. Sasu, M. Lawin, A. Dochhan, J.-P. Elbers, and M. Eiselt, "Advanced optical access technologies for next-generation (5G) mobile networks [invited]," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 12, no. 10, pp. D86–D98, Oct. 2020.

[68] W. Diego, "Evolution toward the next generation radio access network," in *Proc. IFIP Netw. Conf. (Networking)*, 2020, p. 685.

[69] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and learning in O-RAN for data-driven NextG cellular networks," *IEEE Commun. Mag.*, vol. 59, no. 10, pp. 21–27, Oct. 2021.

[70] E. Zeydan, J. Mangues-Bafalluy, J. Baranda, M. Requena, and Y. Turk, "Service based virtual RAN architecture for next generation cellular systems," *IEEE Access*, vol. 10, pp. 9455–9470, 2022.

[71] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 146–172, 1st Quart., 2019.

[72] "5G wireless fronthaul requirements in a passive optical network context," ITU-T, Geneva, Switzerland, ITU-T Recommendation Series G.Supplement 66, Jul. 2019.

[73] J. S. Wey, Y. Luo, and T. Pfeiffer, "5G wireless transport in a PON context: An overview," *IEEE Commun. Stand. Mag.*, vol. 4, no. 1, pp. 50–56, Mar. 2020.

[74] C.-Y. Chang *et al.*, "Slice orchestration for multi-service disaggregated ultra-dense RANs," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 70–77, Aug. 2018.

[75] *Technical Specification Group Radio Access Network; NR; NR and NG-RAN Overall Description; Stage 2. Release 16, V16.1.0*, 3GPP Standard TS 38.300, Mar. 2020.

[76] A. Khan and A. Jamalipour, "Downlink coverage performance of a relay cellular network considering non-uniform user distribution," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2014, pp. 2260–2264.

[77] F. Z. Yousaf, V. Sciancalepore, M. Liebsch, and X. Costa-Perez, "MANOaaS: A multi-tenant NFV MANO for 5G network slices," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 103–109, May 2019.

[78] P. Trakadas *et al.*, "A cost-efficient 5G non-public network architectural approach: Key concepts and enablers, building blocks and potential use cases," *Sensors*, vol. 21, no. 16, p. 5578, 2021.

[79] J.-B. Wang *et al.*, "A machine learning framework for resource allocation assisted by cloud computing," *IEEE Netw.*, vol. 32, no. 2, pp. 144–151, Mar./Apr. 2018.

[80] A. Antonopoulos, "Bankruptcy problem in network sharing: Fundamentals, applications and challenges," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 81–87, Aug. 2020.

[81] D. M. Gutierrez-Estevez *et al.*, "Artificial intelligence for elastic management and orchestration of 5G networks," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 134–141, Oct. 2019.

[82] B. Brik, K. Boutiba, and A. Ksentini, "Deep learning for B5G open radio access network: Evolution, survey, case studies, and challenges," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 228–250, 2022.

[83] A. U. Rehman, R. L. Aguiar, and J. P. Barraca, "Network functions virtualization: The long road to commercial deployments," *IEEE Access*, vol. 7, pp. 60439–60464, 2019.

[84] J. van de Belt, H. Ahmadi, and L. E. Doyle, "Defining and surveying wireless link virtualization and wireless network virtualization," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1603–1627, 3rd Quart., 2017.

[85] *Containers on Virtual Machines or Bare Metal? Deploying and Securely Managing Containerized Applications at Scale*, VMWare, Palo Alto, CA, USA, Dec. 2018.

[86] Y. Wang *et al.*, "Network management and orchestration using artificial intelligence: Overview of ETSI ENI," *IEEE Commun. Stand. Mag.*, vol. 2, no. 4, pp. 58–65, Dec. 2018.

[87] D. Gligoroski and K. Kralevska, "Expanded combinatorial designs as tool to model network slicing in 5G," *IEEE Access*, vol. 7, pp. 54879–54887, 2019.

[88] P. Merle, A. N. Sylla, M. Ouzzif, F. Klamm, and K. Guillouard, "A lightweight toolchain to validate, visualize, analyze, and deploy ETSI NFV TopologiesBehaviors," in *Proc. IEEE Conf. Netw. Softw. (NetSoft)*, 2019, pp. 260–262.

[89] "Network functions virtualisation (NFV); Terminology for main concepts in NFV, V1.3.1," ETSI, Sophia Antipolis, France, ETSI Rep. GS NFV 003, Jan. 2018.

[90] F. B. Lopes, G. L. Nazar, and A. E. Schaeffer-Filho, "VNFAccel: An FPGA-based platform for modular VNF components acceleration," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manag. (IM)*, 2021, pp. 250–258.

[91] M. Pattaranantakul, Y. Tseng, R. He, Z. Zhang, and A. Meddahi, "A first step towards security extension for NFV orchestrator," in *Proc. ACM Int. Workshop Security Softw. Defined Netw. Netw. Funct. Virtualization*, New York, NY, USA, 2017, pp. 25–30.

[92] R. Mijumbi, S. Hasija, S. Davy, A. Davy, B. Jennings, and R. Boutaba, "Topology-aware prediction of virtual network function resource requirements," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 1, pp. 106–120, Mar. 2017.

[93] "Network functions virtualisation (NFV) release 3; NFV security; report on use cases and technical approaches for multi-layer host administration, v1.2.1," ETSI, Sophia Antipolis, France, ETSI Rep. GR NFV-SEC 009, Jan. 2017.

[94] V. Stocker, G. Smaragdakis, and W. Lehr, "The state of network neutrality regulation," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 50, no. 1, pp. 45–59, 2020.

[95] M. Otokura, K. Leibnitz, Y. Koizumi, D. Kominami, T. Shimokawa, and M. Murata, "Evolvable virtual network function placement method: Mechanism and performance evaluation," *IEEE Trans. Netw. Service Manag.*, vol. 16, no. 1, pp. 27–40, Mar. 2019.

[96] J. Cao, Y. Zhang, W. An, X. Chen, J. Sun, and Y. Han, "VNF-FG design and VNF placement for 5G mobile networks," *Sci. China Inf. Sci.*, vol. 60, no. 4, 2017, Art. no. 40302.

[97] F. Alvarez *et al.*, "An edge-to-cloud virtualized multimedia service platform for 5G networks," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 369–380, Jun. 2019.

[98] E. Datsika, A. Antonopoulos, N. Zorba, and C. Verikoukis, "Software defined network service chaining for OTT service providers in 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 124–131, Nov. 2017.

[99] V. A. Cunha *et al.*, "An SFC-enabled approach for processing SSL/TLS encrypted traffic in future enterprise networks," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, 2018, pp. 1013–1019.

[100] T. A. Khan, K. Abbas, A. Muhammad, A. Rafiq, and W.-C. Song, "GAN and DRL based intent translation and deep fake configuration generation for optimization," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, 2020, pp. 347–352.
[101] I. Sarrigiannis, K. Ramantas, E. Kartsakli, P.-V. Mekikis, A. Antonopoulos, and C. Verikoukis, "Online VNF lifecycle management in an MEC-enabled 5G IoT architecture," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4183–4194, May 2020.
[102] J. Cao, Y. Zhang, W. An, X. Chen, Y. Han, and J. Sun, "VNF placement in hybrid NFV environment: Modeling and genetic algorithms," in *Proc. IEEE 22nd Int. Conf. Parallel Distrib. Syst. (ICPADS)*, 2016, pp. 769–777.
[103] "Network functions virtualisation (NFV); infrastructure overview, v1.1.1," ETSI, Sophia Antipolis, France, ETSI Rep. GS NFV-INF 001, Jan. 2015.
[104] D. Camps-Mur *et al.*, "5G-XHaul: A novel wireless-optical SDN transport network to support joint 5G backhaul and fronthaul services," *IEEE Commun. Mag.*, vol. 57, no. 7, pp. 99–105, Jul. 2019.
[105] J. Xia, Z. Cai, and M. Xu, "Optimized virtual network functions migration for NFV," in *Proc. IEEE 22nd Int. Conf. Parallel Distrib. Syst. (ICPADS)*, 2016, pp. 340–346.
[106] "Network functions virtualisation (NFV); infrastructure; compute domain, v1.1.1," ETSI, Sophia Antipolis, France, ETSI Rep. GS NFV-INF 003, Dec. 2014.
[107] J. Zhang, L. Cui, P. Li, X. Liu, and G. Wang, "Towards virtual machine image management for persistent memory," in *Proc. 35th Symp. Mass Storage Syst. Technol. (MSST)*, 2019, pp. 116–125.
[108] C. Xu, H. Wang, R. Shea, F. Wang, and J. Liu, "On multiple virtual NICs in cloud computing: Performance bottleneck and enhancement," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2417–2427, Sep. 2018.
[109] Y. Nakajima, H. Masutani, and H. Takahashi, "High-performance vNIC framework for hypervisor-based NFV with userspace vSwitch," in *Proc. 4th Eur. Workshop Softw. Defined Netw.*, 2015, pp. 43–48.
[110] B. Tak, C. Tang, R. N. Chang, and E. Seo, "Block-level storage caching for hypervisor-based cloud nodes," *IEEE Access*, vol. 9, pp. 88724–88736, 2021.
[111] "Network functions virtualisation (NFV); infrastructure; hypervisor domain, v1.1.1," ETSI, Sophia Antipolis, France, ETSI Rep. GS NFV-INF 004, Jan. 2015.
[112] O. Osanaiye, S. Chen, Z. Yan, R. Lu, K.-K. R. Choo, and M. Dlodlo, "From cloud to fog computing: A review and a conceptual live VM migration framework," *IEEE Access*, vol. 5, pp. 8284–8300, 2017.
[113] "Network functions virtualisation (NFV); infrastructure; network domain, v1.1.1," ETSI, Sophia Antipolis, France, ETSI Rep. GS NFV-INF 005, Dec. 2015.
[114] J. Xu, J. Tang, K. Kwiat, W. Zhang, and G. Xue, "Survivable virtual infrastructure mapping in virtualized data centers," in *Proc. IEEE 5th Int. Conf. Cloud Comput.*, 2012, pp. 196–203.
[115] H. D. Chantre and N. L. S. da Fonseca, "Redundant placement of virtualized network functions for LTE evolved multimedia broadcast multicast services," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2017, pp. 1–7.
[116] M. Maule, J. Vardakas, and C. Verikoukis, "5G RAN slicing: Dynamic single tenant radio resource orchestration for eMBB traffic within a multi-slice scenario," *IEEE Commun. Mag.*, vol. 59, no. 3, pp. 110–116, Mar. 2021.
[117] A. Chagdali, S. E. Elayoubi, and A. M. Masucci, "Slice function placement impact on the performance of URLLC with multi-connectivity," *Computers*, vol. 10, no. 5, pp. 1–18, 2021.
[118] "Network functions virtualisation (NFV); management and orchestration; report on management and orchestration framework, v1.2.1," ETSI, Sophia Antipolis, France, ETSI Rep. GR NFV-MAN 001, Dec. 2021.
[119] A. Alleg, T. Ahmed, M. Mosbah, and R. Boutaba, "Joint diversity and redundancy for resilient service chain provisioning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 7, pp. 1490–1504, Jul. 2020.
[120] H.-G. Kim, D.-Y. Lee, S.-Y. Jeong, H. Choi, J.-H. Yoo, and J. W.-K. Hong, "Machine learning-based method for prediction of virtual network function resource demands," in *Proc. IEEE Conf. Netw. Softw. (NetSoft)*, 2019, pp. 405–413.

MOHAMMAD ASIF HABIBI received the B.Sc. degree in telecommunication engineering from Kabul University, Afghanistan, in 2011, and the M.Sc. degree in systems engineering and informatics from the Czech University of Life Sciences, Czech Republic, in 2016. He is currently pursuing the Ph.D. degree with the Division of Wireless Communications and Radio Navigation, Technische Universität Kaiserslautern, Germany, where he has been working as a Research Fellow since January 2017. From 2011 to 2014, he worked as a Radio Access Network Engineer with HUAWEI. His main research interests include network slicing, network function virtualization, resource allocation, machine learning, and radio access network architecture.

FAQIR ZARRAR YOUSAF (Member, IEEE) received the M.Sc. degree in telecommunication and computers from George Washington University, USA, in 2001, and the Ph.D. degree from the Dortmund University of Technology, Germany, in 2010. He is a Senior Researcher with NEC Laboratories Europe, Germany. He has a total of 17 granted patents, and his research work has been widely published in several peer-reviewed journals, conferences, and book chapters. His current research interest is in NFV/SDN in the context of 5G/6G networks. He is actively involved in the ETSI NFV standards organization, where he is currently serving as the Chair for the Solutions WG and is also a Rapporteur of multiple work items.

HANS D. SCHOTTEN (Member, IEEE) received the Diploma and Ph.D. degrees in electrical engineering from the Aachen University of Technology, Germany, in 1990 and 1997, respectively. Since August 2007, he has been a Full Professor and the Head of the Division of Wireless Communications and Radio Navigation, Technische Universität Kaiserslautern. Since 2012, he has been the Scientific Director with the German Research Center for Artificial Intelligence, heading the Intelligent Networks Department. He was a Senior Researcher, the Project Manager, and the Head of the Research Groups, Aachen University of Technology, Ericsson Corporate Research, and Qualcomm Corporate R&D. During his time with Qualcomm, he has also been the Director of Technical Standards and the Coordinator of Qualcomm's activities in European Research Programs.