Towards zero-touch management and orchestration of massive deployment of network slices in 6G

Hatim Chergui ‡, Adlen Ksentini*, Luis Blanco ‡ and Christos Verikoukis§
*EURECOM, Sophia Antipolis, France
‡CTTC, Barcelona, Spain
§University of Patras, Greece
Email: * adlen.ksentini@eurecom.fr, ‡ firstname.lastname@cttc.es, § cveri@ceid.upatras.gr

***Abstract*—6G systems are expected to serve a massive number of extremely heterogeneous Network Slices that cross multiple technological domains (i.e., RAN, Edge, Cloud, and Core), posing significant challenges to classical centralized management and orchestration approaches in terms of scalability and sustainability. Within this context, distributed and intelligent management and orchestration system is mandatory. This paper proposes a novel framework featuring a distributed and AI-driven management and orchestration system for a massive deployment of network slices in 6G. The proposed framework is compliant of both ETSI standards focusing on autonomous and intelligent network management and orchestration, i.e., Zero touch Service Management (ZSM) and Experimental Networked Intelligent (ENI), leveraging their visions to enable autonomous as well as a scalable management and orchestration of network slices and their dedicated resources.**

I. Introduction

Although 5G has not shown all its potential, 6G foundations have already been devised by the research community. Besides supporting 5G services composed of end-user as well as vertical industry applications, which already constitute a challenge by reporting to 4G services, 6G expects other services that introduce further requirements in terms of latency, reliability, and data rate. 5G has anticipated this evolution by relying on the concept of network slicing, which leverages the flexibility provided by network softwarization to build network instances (virtual networks) tailored to the application or network service. Building on network slicing, 6G will support diverse application and service requirements using the same physical infrastructure. Accordingly, the increased number of services available in 6G will lead to a situation where a massive number of coexisting network slices, with different performance requirements, functionality, and timespans, run in parallel. This puts significant strain on the management and orchestration system that traditional centralized designs, as in Cloud Computing and Network Function Virtualization (NFV) [1], fail to cope with. Indeed, a network slice (or end-to-end network slice) is composed of sub-slices (virtual or physical resources) that belong to different technological domains, i.e., Radio Access Network (RAN), Core Network (CN), Cloud and Edge computing domains. Each technological domain uses its own tools to orchestrate and manage the resources, which complicate the overall management process; hence a centralized solution cannot be envisioned. Moreover, the high number of managed objects involved in the Life-Cycle-Management (LCM) of end-to-end network slices calls for autonomic and self-optimized management and orchestration mechanisms to reduce human intervention to the minimum, hence decreasing the reaction time to sensitive service degradation and avoiding human errors. To overcome these challenges, it is important that the management and orchestration system of network slices combine a hierarchical and fully distributed solution to cope with the heterogeneity of the managed objects of different technological domains, with state-of-the-art Machine Learning (ML) techniques and Artificial Intelligence (AI) algorithms to ensure more autonomy towards zero-touch management.

Recently, ETSI has launched two groups: Zero-touch Service Management (ZSM) [2] and Experiential Networked Intelligence (ENI) [3] aiming to use AI and ML to realize an agile, fully automated management and orchestration of network resources. ETSI ZSM has already issued a reference architecture featuring distributed management and orchestration, but ignoring the management of network slices. Meanwhile, ETSI ENI is more a centralized framework that aims to standardize the different methods and policies to use AI and ML to manage networks. Besides, each group is addressing only a part of the problems related to network slicing management, and no integration of activities is envisioned.

In this paper, we devise a novel decentralized manage-

ment framework that copes with the envisioned massive number and high dynamicity of slices in 5G/6G scenarios, improving both scalability and reaction times of self-management and self-configuration of network slices towards reaching true zero-touch network management. The proposed framework addresses: (1) scalability by relying on a hierarchical and decentralized management system that distributes management functions on several management entities involved in the LCM of network slices, including the network slice itself, which integrates service-level related management functions; (2) zero-touch management by devising a hierarchical closed-control loops that assists the management entities in charge of the LCM of network slices. The proposed framework is compliant with both ZSM and ENI, reaching both groups' objectives within the same architecture. Moreover, the proposed framework is instantiated for two technological domains, Cloud and Radio Access Network (RAN), supporting NFV and O-RAN [4] architectures. Finally, we evaluate our framework through the scenario of slice-level resource prediction under SLA constraints to show the ability: (1) to reduce management overhead; (2) to guarantee SLA using AI/ML federated learning.

The paper is organized as follows. Section II presents the related work. Section III details the proposed management and orchestration framework featuring scalability and zero touch management. Section IV describes the use-case scenario, while Section V presents its performance evaluation. We conclude the paper in VI.

II. Related work

The management of network slices in 6G should be distributed and highly autonomous to handle the high number of managed objects to support the functioning of network slices and guarantee their SLA. Several works in the literature tried to address the challenges related to the management of network slices, either by focusing on scalability or by enabling AI/ML orchestration and management. In [5], the authors proposed a reference architecture featuring a scalable management plan that distributes some management functions inside the network slice (namely in-slice management). However, the proposed work did not explore the usage of AI/ML to leverage the management of network slices. Note that similarly to this work our proposed framework embedded some management functions withing the slice. Work in [6] brings three enabling innovations: (a) Inter-slice control and cross-domain management, to enable the coordination across slices and domains; (b) Experiment-driven optimization to leverage experimental results to design highly performing algorithms; (c) Cloud-enabled protocol stack to gain flexibility in the orchestration of virtualized functions. In [7], the authors explored the usage of AI/ML to leverage the management of network slices adopting the ENI standard concept, which is considered as a highly centralized solution and may not scale well with the high number of network slices. Work in [8] developed an AI-based network management system that provides an intent-based interface for network configuration. In [9], the authors target self-organizing network management mechanisms leveraging the NFV paradigm jointly with AI/ML technologies. But the two mentioned works share the same weakness: the management is still highly centralized and difficult to scale when the number of running slices is high.

On the other hand, standardization groups, such as the ETSI ZSM and ETSI ENI, have been working on using AI and ML to realize an agile, fully automated management and orchestration of network resources. ETSI ZSM Industry Specification Group (ISG) was formed in 2017 to enable full end-to-end autonomous networks capable of self-monitoring, self-healing, and self-optimization without human intervention. However, the ZSM ISG is focused on the definition of generic enablers, closed-loop enhancements, and operations for the next generation of AI-driven autonomous networks, without considering network slicing.

III. Zero-touch AI-enabled distributed management framework

A. Architecture

Figure 1 illustrates the envisioned distributed, zero-touch, and AI-enabled management architecture of network slices. The proposed framework is compliant with both ETSI ZSM and ENI approaches. As stated earlier, the proposed framework is focusing on the LCM of a massive number of network slices and their allocated resources. We assume that an end-to-end network slice is composed of sub-slices run over different technological domains. At least one tenant-specific sub-slice runs the tenant services as Virtual Network Function (VNF), while one or more sub-slices are shared among different end-to-end network slices. An example of a dedicated sub-slice is the tenant's VNF composing the service of the tenant, which are described using the Network Service Descriptor (NSD) and deployed mainly on top of a Cloud or Edge computing infrastructure. A shared sub-slice is a subset of VNF or PNF shared among all the running network slices, e.g., a 5G CN instance or to Central Unit (CU)/Distributed Unit (DU) functions of the RAN. Like ETSI ZSM, we assume that the management and orchestration functions are split

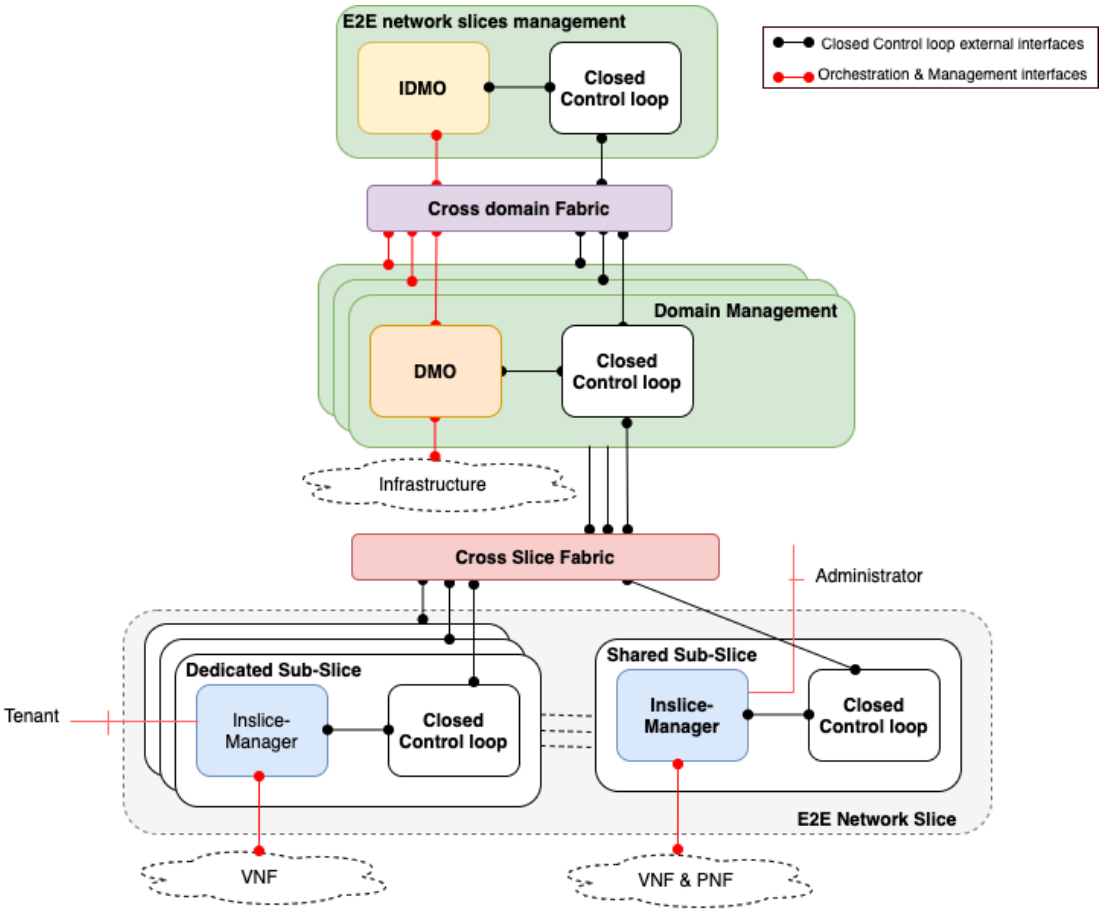


Fig. 1: High-level representation of the proposed architecture.

between the Inter-Domain Management and Orchestration (IDMO) and domain-specific DMO(s). IDMO is a centralized component with full-scope slice management and orchestration decision capabilities. It may take global actions for network-wide, cross-slice, and cross-domain optimizations. For each technological domain, a DMO (e.g., cloud infrastructure, edge, RAN, etc.) may operate its own instances of monitoring and manage specific domain resources in the presence of coexisting slices. Unlike ZSM, the proposed framework argues for further decentralization of the management functions by delegating and enclosing some functions within the running network slices through the InSlice-Manager (ISM). For each slice, the ISM, a logical entity, handles the autonomous management of the slice's functions, i.e., VNF and PNF. ISM can be considered as a slice-level Element Manager (EM), with interfaces to the EMs of the slice's VNFs and PNF.

The IDMO, DMO(s) as well as ISM(s) are leveraged with a closed-control loop that adds intelligence to the management and orchestration functions with AI-based optimization to reach zero-touch service management. The closed-control loops envisioned in this architecture are similar in spirit to the ENI model and correspond to the ENI system that provides the AI/ML to assist external systems for management operations. In the envisioned system, we assume that IDMO, DMO(s), and ISM(s) constitute the ENI assisted-systems, while the closed-control loops provide the AI/ML framework to manage and orchestrate network slices. Indeed, the proposed closed-control loops include the necessary mechanisms and algorithms mainly relying on AI/ML to assist IDMO, MDO(s), and ISM(s) to achieve self-management, self-configuration, and self-adaptation. Besides, the closed-control loops are organised to form a hierarchical control scheme with *fast local* control loops and *slow wider-scope* ones. Two fast local loops are locally run at the DMO(s) and ISM(s), where local decisions on sub-slices can be derived and applied independently and quickly. One wider-scope control-loop that derives decisions at the IDMO level, which has a wide vision on the end-to-end network slice. It should be noted that a limited wide-scope control loop exists at the level of DMOs that involve ISM as well. This closed control loop is involved when service level performances degrade even if ISM has applied local actions. Then, a correlation,

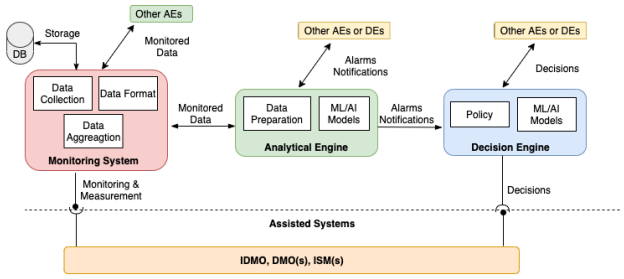


Fig. 2: Closed-control loop components: MS, AE, and DE.

for instance, with MDO resource performances may be needed to solve the issue. Using this hierarchical scheme permits to leverage of time-scale decomposition at different levels of the proposed system, hence limiting the interference among different feedback-based decisions. Moreover, using local data processing and decisions, we (a) minimize the exchange of (big) data between components (e.g., aggressive feature selection) to keep management scalable, and (b) significantly reduce the reaction time of data-driven management decisions that could be handled locally.

Similar in spirit to ZSM, all the entities involved in the management and orchestration process are using the cross domain fabrics and cross slice fabrics to expose and consume APIs. Here we distinguish between the external interfaces used by the closed control loop to collaborate aiming at assisting IDMO, DMOs and ISMs to derive decisions to handle network slices' LCM, and the interfaces between IDMOs and DMOs which are mainly used for orchestrating and managing sub-slices resources.

B. Closed-control loops

One of the critical features of the proposed framework is the closed-control loops that assist IDMO, DMOs, ISM with AI capabilities to reach the zero-touch management objective. Figure 2 shows the components of the Closed-control loop: Monitoring System (MS), Analytical Engine (AE), and Decision Engine (DE). These three elements are already known in the Infrastructure as a Service (IaaS) management process of virtual resources, principally relying on a centralized element (i.e., orchestrator) that runs the three entities. However, in the proposed approach, the three elements are highly distributed among the actors managing and orchestrating network slice components and resources. MS is in charge of monitoring Key Performances Indicators (KPI) and various relevant events from the different components deploying network slices (NFV Infrastructure - NFVI, SDN Controller, RAN, etc.) as well as from the VNFs and PNFs composing a network slice. Note that MS may use a DB to store the collected monitoring data for future analysis. It can be used for instance to understand the long term evolution of a system. MS periodically transmits monitoring information to AE that processes the data and provides the required analysis output to the DE. The latter, using a pre-deployed policy or automatic decision mechanisms, decides on the LCM action to apply; in the case of a VNF a LCM action is to scale up or down the resources (e.g., CPU) or migrate the VNF to another NFVI. These actions are enforced by the DE using interfaces with the components managing the network slices (i.e., IDMO, DMOs, and ISM). AE and DE are highly driven by AI/ML techniques aiming at learning appropriate LCM decisions to consider, according to the state of the infrastructure, the state of network slice components (obtained from AE), and depending on the technological domain where each slice component is deployed (e.g., RAN).

The Closed-control loop assisting IDMO is different from those assisting DMOs and ISMs, as no monitoring information is extracted from IDMO. Therefore, the IDMO is exposing to the closed-control loop only the possible actions to update/upgrade the running end-to-end network slices. These actions will be considered by DE as possible decisions to follow the AE recommendation. Example of such action is to migrate a sub-slice from one infrastructure provider to another one reacting to service degradation or security threats. On the other hand, DMOs' decisions are local to the technological domains. The decisions mainly concern the management of the resources that are used by network slices, such as scale up/down VNFs or update the RAN resources dedicated to a running slice. Finally, the ISM decisions cover mainly service-level configuration. For instance, change the video encoding, update the flight plan of flying drones, etc.

1) Monitoring System (MS): MS role is to collect critical information on the functioning of a system and provides this information, after, for example, aggregation or normalization, to AE, which in turn uses this information to detect and react to network slices' LCM events, such as performance degradation, performance optimization, and security threats. MS interacts with different entities that orchestrate and manage the per technological domain sub-slice, i.e., DMOs. Further, MS interacts, through ISM, with slice-specific VNFs and applications, as well as shared VNF and PNF among network slices. Indeed, we distinguish between information that monitors the state of the infrastructure shared by the running slices and the information that monitors the

VNF of tenants and applications. For the infrastructure monitoring, MS has to interact with DMO(s) to collect information on:

- NFVI, such as computing platforms and hardware;
- PNFs running network functions on dedicated hardware, such as eNB/gNB, router, and User Plan Function (UPF);
- VNFs running common virtualized network functions, such as 5G Core Network (CN) functions or Directory Name Service (DNS).

Regarding function monitoring, MS has to interact with VNFs or applications of the tenants through EM exposed by VNFs or PNFs. EM is a set of API exposed by network functions that allow extraction of information on the state of the application, such as events, alarms, and logs, but also permit to act on the application behavior to update its configuration.

The principal consumer of MS information is AE, which is in charge of triggering the monitoring of needed information from MS. The latter starts the monitoring process by connecting to the appropriate source. Accordingly, MS exposes two types of APIs: control API and data collection API. AE uses the control API to request the KPI to monitor, the periodicity, the duration, etc., while the data collection API is the interface from which data is provided to AE as requested through the control API. The control API also indicates how data is provided, i.e., publish/subscribe, request/response, the data format, etc.

2) Analytical Engine (AE): As opposed to MS, AE does not store, but processes data gathered from the same or lower-level MS or AE and exposes the result to any requester (i.e., DE or other AE) in an on-demand or periodic fashion. AE to AE communication is possible to build a learning model using Federated Learning (FL) techniques. The main functions of AE are: (i) identify performance degradation of a network slice; (ii) optimize the performance of a network slice or the DMO resources; (iii) react to security threats. To this aim, AE subscribes for data types to which it is interested in using the control API exposed by the MS. The data type will be determined according to the logic of the LCM application execution. Then, AE starts receiving the stream of data or uses a request/response mechanism, depending on the purpose of the analysis. AE may adapt the monitoring data rate or stop the precedent request and request for other related monitoring information. AE heavily relies on AI/ML to complete an inference task locally, extract features, and analyse these features and send alerts and notifications to DE. AEs can collaborate to build distributed learning (based on federated learning) models to realize the analysis and notify the DE accordingly. Examples of features extracted and analysed are: prediction of SLA violation, prediction of service migration, prediction of NS Faults, attack Identification, anomaly detection. Note that that an example of predicting SLA through AEs using federated learning is presented in section IV

3) Decision Engine (DE): DE is the decision-making element of the proposed framework. It analyses alerts and notifications from AE(s) and considers a decision to take. The decisions are either derived using a local ML algorithm, based mainly on Reinforcement Learning (RL), or a predefined policy enforced by the Tenant or DMO through Intent, or a combination of both. DE may collect notifications from several AEs of different TDs to consider wide-scope decisions on the end-to-end network slices. DE uses exposed APIs by DMOs to enforce the considered decisions. For local decisions, DE interacts with DMO and ISM, while for global decisions, the DE has to interact with IDMO. Examples of global decisions are energy optimization, block UE connection, while local decisions are: VNF scaling, update RAN resources dedicated to a slice, and service migration.

C. IDMO

IDMO is equivalent to the 3GPP Network Slice Management Function (NSMF) [10] and exposes the Northbound API (NBI) for the OSS/BSS or Consumer Service Management Function (CSMF). IDMO is in charge of the LCM of end-to-end network slices. It has full-scope slice management and orchestration decision capabilities and takes global actions for network-wide, cross-slice, and cross-domain optimizations. The tenant or the slice owner interacts with the OSS/BSS or CSMF to define the network slice to deploy using an already Blueprint to generate a Network Slice Template (NST) that includes attributes and meta-data on the network slice (ex. the start date and end date, slice owner, type of slice, etc.), and information on each sub-slice composing the network slice. For instance, in the case of computing resource (i.e., Cloud or Edge) domain, the NST may include information such as the number of CPUs, memory, and virtualization technology (i.e., VM or containers) to be used. For the RAN domain, resources may be related to the functional split type [11], the MAC scheduler algorithm, the number of Radio resource Blocks (RB), and others. Finally, for the transport domain, resources may include the type of link (bandwidth, latency), number of Virtual Local Area Networks (VLAN)s, front haul link capacity, Virtual Private Network (VPN) links, and QoS. Each technological domain needed resources are enclosed in the NST in the form of a technological

domain-specific descriptor. For instance, for the NFVI domain, the resources are described using a Network Service Descriptor (NSD) that includes the VNF(s) list and their descriptors.

In the proposed framework, the NST also includes management functions to be embeded within the network slice, i.e., the closed control loops (MS, AE, and DE) and ISM template. Upon receiving the slice creation (i.e. NST), IDMO is in charge of selecting the infrastructure provider to run the sub-slice. To this end, it uses a resource broker that relies on a specific algorithm to select the appropriate infrastructure provider for each technological domain where to deploy sub-slices composing the end-to-end network slice. After that, the NST is split into a technological-specific template and forwarded to the DMO of the selected infrastructure providers.

D. DMO

Each technological domain is managed and orchestrated by its own entity, Domain Management Orchestrator (DMO), which is equivalent in 3GPP to the Network Sub-Slice Management Function (NSSMF)[10]. Depending on the technological domain, a NSSMF may correspond to NFVO for Cloud/Edge, RANO for RAN, and Software Defined Networking (SDN) controller for the case of the transport network.

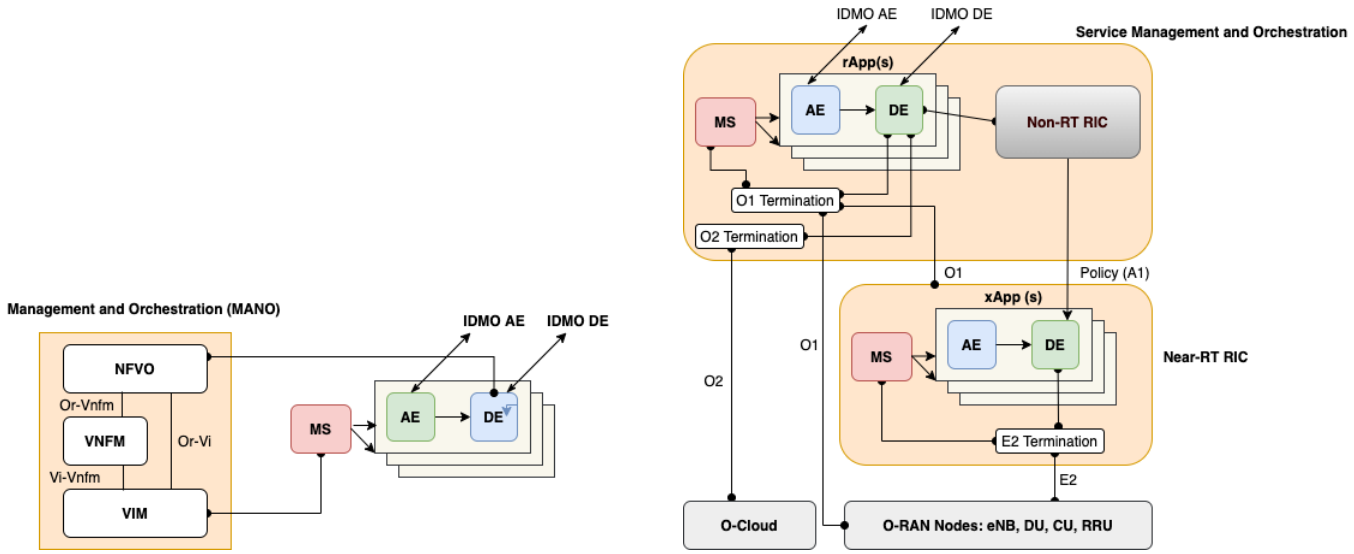
Figures 3a and 3b illustrate the mapping of DMO with its closed-control loops for two different technological domains, cloud and RAN, respectively. For the cloud domain we use the well-known NFV architecture, while for the RAN domain we consider the emerging O-RAN architecture.

In the case of the cloud and edge (Figure 3a), MDO corresponds to the MANO entity, constituted by the NFV Orchestrator (NFVO), VNF Manager (VNFM), and Virtual Infrastructure Manager (VIM). Monitoring information is collected by MS from VIM covering mainly infrastructure-level metrics: such as CPU and memory usage, consumed throughput, and packets per second; organised per VNF or aggregated per sub-slice. Besides, the NFVO may expose available management actions that can be performed on a cloud or edge sub-slice, which DE can use. We can mention VNF scale-in or out, VNF migration, block incoming or outgoing VNF, instantiate a new VNF, etc. Typically, all the actions are related to VNFs LCM. It should be noted that DE assisting MANO may refer to the IDMO's DE, if a local decision is not enough to resolve an issue. For example, if no more CPU is available at the infrastructure level or need to migrate one VNF between two VIMs, the local DE may delegate the decision to the IDMO's DE that may request more details from IDMO's AE or use a local policy to select a new VIM to serve the sub-slice, using a multi-domain placement algorithm [1]. Finally, MS is common for all running AE/DE; it collects monitoring data, format, and aggregate data to be consumed by subscribed AEs.

Regarding the RAN domain (Figure 3b), the MDO encloses Service Management and Orchestration (SMO) and the Near Real-Time RAN Intelligent Controller (Near-RT RIC). While, the closed-control loops are run as rApps (i.e., applications that run Non-RT RIC) in SMO for everything related to non-RT management loops, such as Fault-management, the configuration of CU/DU, and LCM of xApps (i.e. applications that runs Near RT-RIC); and xApps for everything related to Near-Real time management functions, such as MAC scheduling, Mobility management, Radio resources management, etc. It should be noted that according to O-RAN the O-Nodes correspond to RAN functions, either run as a monolithic block (gNB), as PNFs, or disaggregated functions: CU, DUs, and Radio Units (RU)s. CU and DU may run as VNFs on top of a cloud or edge infrastructure (noted as O-Cloud), while RU is a physical component that runs as a PNF.

Closed-control loops running in the SDM collect monitoring data on O-RAN nodes in addition to Near-RT RIC components and O-Cloud, using O1 and O2 interfaces, respectively. O1 allows collecting data on xApps performances, logging, and monitoring on the status of Near-RT RIC internals components, radio information extracted from CU/DU/RU. On the other hand, O2 provides cloud-oriented information similar to what VIM provides, focusing only on the VNF running CU and DU, such as consumed CPU and memory, exchanged traffic, etc. MS in charge of collecting the monitoring data using these two interfaces, while AEs will subscribe to the data of interest. DEs may enforce the derived decisions using either O1 and O2. The actions are related to the configuration of CU/DU (for instance, turn-off DU to save energy or change the configuration of the Time Duplex Division (TDD) pattern) or on the scale-up/down of VNFs running CU/DU. Moreover, DE may use the A1 interface, via the Non-RT RIC to push a new policy toward the running xApps, for instance, to increase the RB dedicated to a slice, or change the 5G New Radio numerology [12], or update the MAC scheduler of a network slice.

Closed-control loops running in Near-RT RIC are concerned with real-time management of RAN. Only one interface is available to collect and act on the RAN functions (i.e., CU, DU, and RU), E2. This interface



(a) Cloud domain. (b) RAN domain

Fig. 3: MDO of cloud and RAN technological domains.

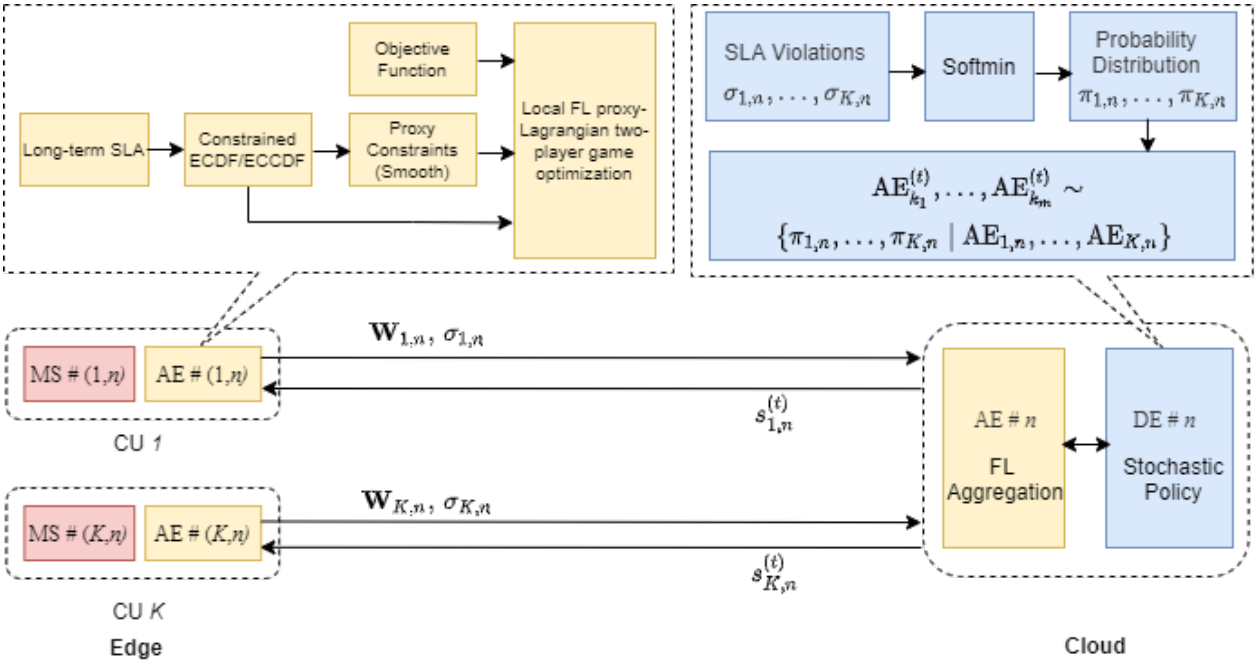


Fig. 4: Cross-domain stochastic policy for low Slice-Level SLA violation and overhead

will allow MS to collect monitoring information on the radio side, covering User Equipment (UE) and Cells information, like the used RB per sub-slice, the number of attached users per cell, per UE reported channel quality, etc. The DE can, for instance, act by increasing the number of RBs per sub-slice, trigger handover, change the 5G NR numerology to be used by a UE, the scheduled UEs for the next period, shift the TDD pattern, etc.

E. ISM

The Inslice-Manager is the finest management entity that handles fault management, configuration management, and performance management of the service deployed on top of the sub-slice. We distinguish between ISM of tenant-specific sub-slice that runs the tenant network service and the shared sub-slice managed by the network operator. Both ISMs are assisted by a closed-control loop to handle service levels related performances. MS monitors the service level KPI, such

as user-perceived Quality of Experience (QoS), number of users served by a VNF, logs of the applications, etc. AE analyses and extracts features related to the service level performances, such as services' response time degradation, user-perceived QoE degradation. DE derives decisions mainly at the service level, such as the change of VNF configuration, change the video encoding, block user traffic, etc. To this end, DE passes through the ISM, which will enforce the decision using the VNF's EM.

Regarding the shared sub-slice, for instance, 5G CN instances, which usually belong to the network operator, the ISM is concerned mainly with the service level performances of the shared VNFs and PNFs. The closed-control loop may supervise the performance of the Authentication and Mobility Function (AMF) by detecting performance degradation due to overloaded received attach or to DDoS attacks. MS may monitor information such as the number of attaches of UEs seen by the AMF, the number of treated packets by the UPF, etc. AE predicts and detects issues, such as service degradation (increasing of the attach duration or malicious traffic). Therefore, DE decisions can change the configuration of the AMF, block UEs attaches in case of security threats, drop packets at the UPF, etc.

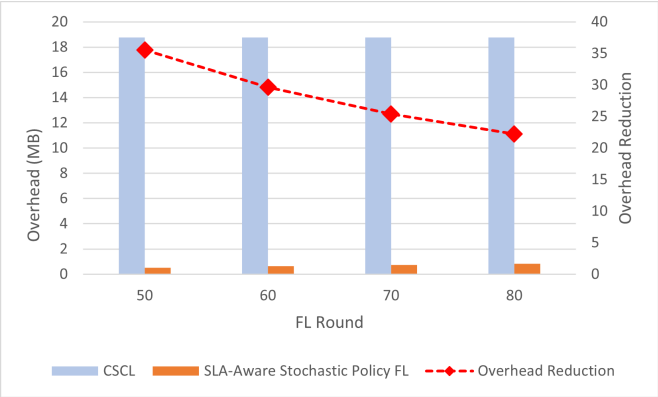
It should be noted that one of the decisions of DE (of both ISM) is to send an alert to the DMO where the VNFs are run when local decisions are not sufficient. This will allow investigating further the reasons leading to bad service performances and may lead the DMO DE to take local decisions, for instance, by doing a scale up of resources of the VNFs. If DMO's local decisions are not efficient, the DMO DE may refer to the upper layer DE (at IDMO) to consider other solutions at the end-to-end level, such as selecting another technological domain provider.

IV. Use-case scenario

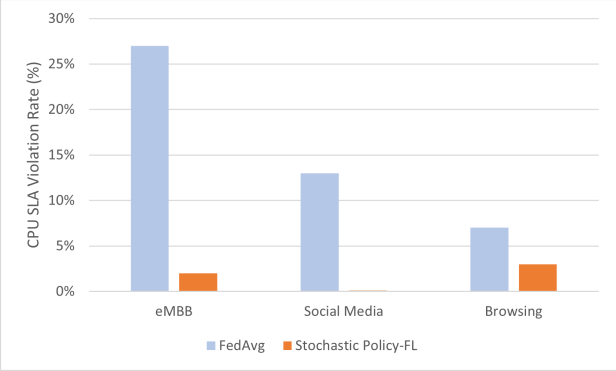
To showcase both the scalability and zero-touch management capabilities of the proposed management framework, the use-case of slice-level resource prediction under SLA constraints is considered, where the objective is to respect the SLA violation rate of each slice while dramatically minimizing the management overhead. In this regard, Figure 4 details the deployment of the use-case on top of the architecture where, under the central unit (CU)-distributed unit (DU) functional split, K CUs are running as VNFs at the Edge, and including co-located MS and AE which are instantiated per slice. To avoid exchanging raw monitoring data with the cloud domain, this SLA-constrained resource prediction is performed locally by each AE (k, n). It consists of learning a resource provisioning regression model under long-term SLA constraints and given some space-time varying input features such as slice traffic and radio condition—that depend on the CUs locations at RAN and slice type. A typical SLA between slice n tenant and the network slice provider would consist on imposing an upper-bound γ_n on the probability that a slice resource usage exceeds an interval $[\alpha_n, \beta_n]$, which translates into learning the AE local resource prediction model under empirical cumulative density function (ECDF) and complementary CDF (ECCDF) constraints which are solved via proxy-Lagrangian two-player game [13]. Since the local MSs datasets are not exhaustive, the local AEs participate in a federated learning (FL) task to improve their prediction, where only their slice n models weights and achieved SLA violation rates $\{\sigma_{k,n}\}_{k=1}^{K}$ are reported to the end-to-end AE and DE, respectively, which are both located at IDMO. At each FL round, the end-to-end AE averages the received weights and broadcasts the result to the decentralized AEs to update their local models. To minimize the computation and management data, the end-to-end DE (at IDMO) implements a stochastic policy to select the local AEs that can take part in the FL optimization. To this end, it uses the received SLA metrics to generate a probability distribution over the AEs using the `softmin` activation layer. This means that AEs whose models achieve a low SLA violation rate are assigned a higher probability. Specifically, at each FL round t, the end-to-end DE chooses $m < K$ local AEs to participate in the FL task based on the probability distribution $\{\pi_{1,n}, \ldots, \pi_{K,n}\}$, and feeds back an activation bit $s_{k,n}^{(t)} \in \{0, 1\}$ to inform the AEs. In this case, only the selected AEs would train and send their weights to the end-to-end AE for aggregation, but the generated global model is broadcast to all AEs thereafter. This procedure allows to orient the FL training to the models yielding a low SLA violation and ensures that, by the convergence round (i.e., in the long-term), the AEs would have stochastically taken part in the FL task according to the initial probability distribution, while avoiding to concurrently involve all the AEs in the training at each round. Hence, this strategy further minimizes the system overhead and computation load as well as paves the way to accept and process a massive number of slices.

V. Performance Evaluation

To assess the above use-case, a scenario with K CUs is considered, where due to the heterogeneous space-time traffic patterns and radio conditions at various transmission/reception points (TRPs) over the RAN,



(a) Overhead reduction.



(b) CPU load average violation rates with $\alpha = [0, 0, 0]$, $\beta = [4, 7, 10]$ % and $\gamma = [0.01, 0.01, 0.01]$.

Fig. 5: Evaluation results.

the corresponding MSs hold non-independent identically distributed (NIID) datasets of size $D_{k,n} = 1000$ that have been collected in our evaluation from a live LTE-advanced network. The datasets include, as input features, the hourly traffics of the main over-the-top (OTT) applications, channel quality indicator (CQI) and MIMO full-rank usage, while the considered supervised output is the CPU load. The assumed slices are:

- enhanced Mobile BroadBand (eMBB): NetFlix, Youtube and Facebook high and ultra-definition videos,
- Social Media: Facebook, Facebook Messages, Whatsapp and Instagram,
- Browsing: Apple, HTTP and QUIC,

By coding the samples of all the datastes, the AEs' models weights as well as the SLA violation metrics in 32 bits format, we quantify the overhead induced by both the proposed stochastic policy-FL and a fully centralized SLA-constrained learning (CSCL) baseline [14] as depicted in Figure 5a where $m = 100$. Intuitively, since only the AE models' weights are exchanged in the FL setup instead of the raw dataset, as is the case in CSCL, the overhead is expected to be significantly reduced. Concretely, upon the convergence point of the stochastic policy-FL, i.e., round 50 more than $\times 30$ overhead reduction is obtained, with a minimum of $\times 22$ reduction at round 80. While also dramatically reducing the CPU SLA violation rate compared to the FedAvg unconstrained algorithm [15] as showcased by Figure 5b.

VI. Conclusion

In this paper, we introduced a novel distributed management and orchestration framework that addresses the challenge of handling a massive number of network slices as envisioned in 6G. The proposed framework relies on a hierarchical AI-driven close control loop to assist management entities in handling autonomously and efficiently network slices LCM. A representative use-case scenario has been introduced to demonstrate the usage of the framework to achieve zero-touch management to guarantee SLA. The proposed framework is compliant with ETSI ZSM and ENI, and has been mapped to two major architecture to manage cloud and RAN resources, namely NFV and O-RAN.

VII. Acknowledgement

This work was partially supported by the European Union's Horizon 2020 Research and Innovation Program under the MonB5G project (Grant No. 871780).

References

[1] N. Toumi et al., "On Using Physical programming for Multi-Domain SFC Placement with Limited Visibility", in IEEE Transactions on Cloud Computing, 2020.
[2] ETSI Zero touch network & Service Management (ZSM), www.etsi.org/committee/zsm [Accessed in May, 2021].
[3] ETSI, "Experiential Networked Intelligence (ENI)", [Online]. Available: www.etsi.org/committee-activity/eni (accessed: Jan. 19, 2021).
[4] ORAN Alliance, 2020b. "Operator defined next generation ran architecture and interfaces". URL:https://www.o-ran.org/
[5] S. Kukliński et al.,"A reference architecture for network slicing," 2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft), 2018, pp. 217-221.
[6] Bulakci, Ö., et al., "Overall 5G-MoNArch architecture and implications for resource elasticity, in IEEE European Conference onNetworks and Communications (EuCNC) 2018, June 18-21, Ljubljana,Slovenia
[7] D. Gutierrez-Estevez, et al. "Artificial Intelligence for Elastic Management and Orchestration of 5G Networks." IEEE Wireless Communications Magazine, 2019.
[8] D. Camps-Mur et al. "5G-CLARITY: Integrating 5GNR, WiFi and LiFi in Private Networks with Slicing Support", IEEE European Conference onNetworks and Communications (EuCNC), June 2020.

[9] L. Baldini et al., "SliceNet Control Plane for 5G Network Slicing in Evolving Future Networks," 2019 IEEE Conference on Network Softwarization (NetSoft), 2019, pp. 450-457.
[10] 3GPP, "Management of network slicing in mobile networks; Concepts, use cases and requirements", 3GPP TS 28.530, ver. 17.1.0, Apr. 2021.
[11] C.Y. Chan et al."Slice Orchestration for Multi-Service Disaggregated Ultra-Dense RANs.", IEEE Communications Magazine, Vol. 56, Issue 8: 70-77 (2018).
[12] K. Boutiba et al., "NRflex: Enforcing Network Slicing in 5G New Radio", to appear in Elsevier's Computer Communications journal.
[13] A. Cotter *et al.,* "Two-Player Games for Efficient Non-Convex Constrained Optimization," [Online]. Available: arxiv.org/abs/1804.06500v2.
[14] H. Chergui and C. Verikoukis, "Offline SLA-Constrained Deep Learning for 5G Networks Reliable and Dynamic End-to-End Slicing," in *IEEE Journal on Selected Areas in Communications,* vol. 38, no. 2, pp. 350-360, Feb. 2020.
[15] H.-B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data.", in *the 20th International Conference on Artificial Intelligence and Statistics (AISTATS'2017).*

BIOGRAPHIES

Hatim Chergui is the project manager of the H2020 MonB5G European project and a Post-Doctoral researcher at CTTC, Spain. He served as an expert at both INWI and Huawei Technologies, Morocco. He was the recipient of the IEEE ICC 2020 Best Paper Award. He is an Associate Editor at IEEE Networking Letters.

Adlen Ksentini is an IEEE COMSOC distinguished lecturer. He obtained his Ph.D. degree in computer science from the University of Cergy-Pontoise in 2005. Since March 2016, he is a professor in the Communication Systems Department of EURECOM. He has been working on several EU projects on 5G, Network Slicing, and MEC.

Luis Blanco Luis Blanco received the MSc and PhD degrees in Telecommunications engineering from the Polytechnic University of Catalonia (UPC), Spain. He holds a degree in Research on Information and Communications Technologies and in Data Science and Big Data. His current research interests include AI/ML for sustainable B5G/6G wireless networks, M2M/IoT over satellite systems and ultra-dense wireless networks.

Christos Verikoukis is an Associate Professor (Tenure Track) in the University of Patras, Department of Computer Engineering and Informatics. He got his Phd from Technical University of Catalonia in 2004. He has published more than 140 Journal and 220 Conference papers and 3 Books. He is currently the IEEE ComSoc EMEA Director, an IEEE ComSoc BoG member and a GITC member-at-large.