Received: 15 September 2021 | Revised: 22 March 2022 | Accepted: 24 June 2022

DOI: 10.1002/ett.4598

RESEARCH ARTICLE

WILEY

A unified service-based capability exposure framework for closed-loop network automation

Marco Gramaglia[1] | **Marton Kajo**[2] | **Christian Mannweiler**[3] | **Ömer Bulakci**[3] | **Qing Wei**[4]

[1]Department of Telematic Engineering, Universidad Carlos III od Madrid, Leganes, Spain

[2]Department of Informatics, Technische Universität München, TU München, Germany

[3]Nokia Bell Labs, Germany

[4]Huawei German Research Center, Germany

Correspondence
Marco Gramaglia, Universidad Carlos III de Madrid, Avenida Universidad 30, 28911 Leganes, Spain.
Email: mgramagl@it.uc3m.es

Funding information
European Commission; H2020 5G-MoNArch, Grant/Award Number: 761445; H2020 5G-TOURS, Grant/Award Number: 856950; Spanish Ministry of Economic Affairs and Digital Transformation and the European Union, Grant/Award Numbers: 6G-CLARION-NFD, 6G-CLARION-OR, 6G-CLARION-SI, 6G-CLARION-OE

Abstract

The ongoing quest for the tight integration of network operation and the network service provisioning initiated with the introduction of 5G often clashes with the capacity of current network architectures to provide means for such integration. Owing to the traditional design of mobile networks, which barely required a tight interaction, network elements offer capabilities for their continuous optimization just within their domain (eg, access, or core), allowing for a "silo-style" automation that falls short when aiming at closed-loop automation that embraces all the actors involved in the network, from network functions up to the service-provider network functions. To this end, in this article, we make the case for the network-wide capability exposure framework for closed-loop automation by (i) defining the different entities that shall expose capabilities, and (ii) discussing why the state of the art solutions are not enough to support this vision. Our proposed architecture, which relies on registration and discovery, and exposure functions, allows for enhanced use cases that are currently not possible with state of the art solution. We prove the feasibility of our solution by implementing it in a real-world testbed, employing Artificial Intelligence algorithms to close the loop for the management of the radio access network.

1 | INTRODUCTION

The continuous development toward more flexible networks[1] makes softwarization a key enabling technology that has also impacted standardization during the last few years; starting from the introduction of software-defined networking (SDN) and followed by the network function virtualization (NFV) concepts.[2] Recently, many network-related standards adopted service-based architecture (SBA)[3] principles that reference service-oriented architecture (SOA) paradigms.[4] Here, a network function (NF) can flexibly communicate with other NFs to consume the provided services, thus overcoming the limitations of a reference-point-based interaction, where an interface is only defined between two NFs.

Moreover, this has opened up new opportunities,[5] for example, the recently developed network slicing[6] paradigm. On the one hand, the flexibility given by a programmatic approach to network management and control has allowed

The copyright line for this article was changed on 24 August 2022 after original online publication.

the possibility of creating network instances (ie, the network slices), tailored to various applications and service provider needs on the same infrastructure. However, on the other hand, this introduced additional complexity in the management side, due to the dependencies between slices that require different optimal operating points but have to share a common underlying network infrastructure.

1.1 | Data-driven network management

These new paradigms indeed represent a significant development step compared to the rather monolithic structure of legacy mobile networks, which has been basically designed to provide mobile broadband services over one physical network instance. In order to manage such a possibly large number of network instances, big data,[7] and artificial intelligence (AI) techniques have been considered[8] as potential enablers for autonomous management of the network. Features such as auto-scaling, self-optimization, or intent-based networking[9] clearly fit well with a data-driven approach to the network environment, in which elements offer ways to produce and consume data, but also can be configured according to service-related policies in a jointly optimized way. Achieving autonomous network management thus is a challenging task that entails overcoming a number of technical complexities:

- **Data heterogeneity**: Autonomous management shall be applied at all levels of the network, from single NFs to large infrastructure deployments. Hence, the data collected from the network include very different metrics ranging from typical network key performance indicators (KPIs), for example, actual throughput and latency, to general-purpose resource utilization, for example, CPU load. Moreover, the output data generated by different analytics functions have even more heterogeneity.
- **Temporal timescale heterogeneity**: Decisions may be applied at very different temporal scales, depending on the type of element that has to be assisted. A user plane (UP) function may need to be assisted at almost wire speed, while orchestration strategies work on larger time scales (eg, hours).

In addition to the challenges from the scalability point of view (ie, the autonomous management of the network shall be performed on many and potentially large network instances), another fundamental technical barrier relates to the interfaces needed across different network domains. While the term *domain* has been extensively used in the literature to distinguish individual network segments by technology, implementation architecture, or administrative ownership (tenant concept),[10] in this article, we use the term *network domain* to refer to different functionality within and beyond the considered network. The new challenges that the operation of the network poses to all the elements entails a deep revision of how different parts inside and outside the network interact with each other, enabling a much higher granularity that goes beyond the pure ownership of the infrastructure and involves, instead, different network domains, as discussed next.

1.2 | The need for a new exposure paradigm

Traditionally, procedures related to network management, network orchestration, and network control have been developed by different tracks of the standardization activities. Hence, the functions that execute these procedures (ie, OSS functions, element managers, orchestrators, or radio and core NFs) have been designed in a domain-specific and, in some cases, even proprietary manner, with possible optimizations happening only in a "per domain" way, leaving the interaction limited to peer-to-peer reference points within a domain, for example, between control plane (CP) and UP (eg, the S11 interface between MME and S-GW functions as defined in 3GPP 4G evolved packet core[11]).

In such reference-point-based setups, optimizations are either open-loop (ie, no feedback among different modules in the system) or require very expensive human engineering procedures to close the loop. This approach has been deemed as valid within legacy networks, due to their limited number of possible configurations. However, in a 5G environment, this approach clearly falls short. Also, as legacy NFs either have function-specific data acquisition and processing procedures or no procedure at all, simple configurations or rulesets are usually sufficient to achieve optimization goals.

Self-organizing network (SON)[12] functions were the first step toward that goal. Frequently, however, the legacy approach is insufficient for the 5G environment with network slicing. With legacy SON, configurations mainly concern

"physical" parameters (eg, radio transmission power, radio conditions, handover thresholds, cell-individual offsets, and antenna tilt, to only name a few), while per-slice configuration or the self-optimization of the NFV infrastructure is simply out of the scope of existing SON concepts. Still, SON is a first step toward the flexible management of the network without the human in the loop.

In 5G, network slicing, among others, has imposed a more modular design of NFs, allowing these NFs to be shared and re-used across slices in a more fine-grained and targeted manner.[13] That is a single Network Slice Subnet Instance (NSSI) and its constituent NFs may be used across several slices and services (eg, common radio access NFs across slices). Thus, interfaces devoted to the data acquisition and processing from NFs or even to feed and push data to specific AI modules shall be designed. In order to "close the loop" in an automated manner, by adding AI and big data solutions, new interfaces and functionalities are needed:

- **Flexible data exchange across domains**: Different network domains shall be able to exchange information among them using, for example, a publish-subscribe methodology.
- **Reliable and scalable configurations**: Besides producing and consuming the data, the NFs in all domains shall offer ways to allow authorized configuration of their relevant parameters, possibly with different levels of authorization, for example, depending on the enforced resource provisioning scheme. For instance, some service providers may have full configuration capabilities while others may only have limited visibility of configurable parameters.

We remark that the current, state-of-the art network architectures are currently defined in a "silo"-based way: they may have some analytics feature within one domain (such as the notable case of NWDAF in the core), but they are lacking the open exchange among them. Thus the motivation of this work encompasses aspects such as: (i) identifying the possible data exchanges among different domains, classified according to their purpose, (ii) finding an architecture that overcomes this issue, and (iii) exemplify it in a compelling way.

1.3 | Contributions

In this article, we make the case for a novel, network-exposure native, network architecture, that bridges the gaps that are traditionally found across different Standard Developing Organizations (SDOs), which often focus on a specific domain only.

Thus, the contributions of this article can be summarized along three different areas, as follows.

- A thorough review of the state of the art on the topic of network exposure across different domains. Also, we discuss here a novel definition of network domain that goes beyond the one traditionally used in the literature. This is detailed in Section 2.
- A taxonomy of producers and consumers (discussed in Section 3) in the network architecture and the kind of capabilities they can exchange. To the best of our knowledge, this is the first work that tackles a network-wide exposure framework, discussing its fundamentals.
- A new set of network procedures (detailed in Section 4), designed around the exposure concept, that facilitates the introduction of data-driven algorithms for network management and operation.
- Some possible use cases that will benefit from the proposed architecture (in Section 5). Additionally, we provide some feedback on real life experiments on a trial network deployment for one of the use cases, running in the Hamburg Port, Germany.

Hence, this article provides a holistic view on the advantages provided by enabling a native network exposure architecture, joining state of the art discussion with a feasibility study in a real-world scenario. Hence, the role of the deployed system discussed in Section 6 is to provide a feasibility analysis of the overall network exposure system, effectively showcasing one of the provided case studies. In other words, we discuss a real-world implementation of a specific algorithm, with the goal of providing qualitative evidence of the advantages provided by the service based capability exposure framework proposed in this article.

2 | DOMAINS AND CAPABILITIES

A common distinctive aspect for the next generation of mobile networks with respect to legacy ones is their tighter interaction between service provider and network operators. The final goal is achieving a continuum between the end users and the provided services through the network infrastructure, that is tailored for the specificity of the envisioned application. As a result, the network becomes a kind of "commodity" for the service provider, that has to be managed as many other infrastructure deployments.

This totally new paradigm, enabled by network softwarization and programmability, also allows to employ data-driven solutions for steering the operation of the network. However, like any other algorithm based on Artificial Intelligence, the availability of input data (used, eg, to train models) and the possibility of enforcing the decisions stemming from these models into the network becomes fundamental.

However, the state of the art solutions for the network architecture only provide a limited availability of data across different domains in the network. With domain we do not refer to administrative domains (ie, network deployments that belong to different operators), but rather to different elements of the network that perform different networking tasks (eg, network control, network management, or network deployment).

In the following, we divide the network elements into four domains, as depicted in Figure 1: network functions, management, orchestration, and service providers. For each of them, detailed in Section 2.1, we discuss the different state-of-the-art solutions for the production of network exposure data.

We further structure the discussion into which capabilities may be exposed (enumerated in Section 2.2, and the possible consumers analyzed in Section 3.

2.1 | The domains

With the arrival of the network softwarization concept, different functionalities in the network adopted a software-driven approach. Among the most important examples, we can list the Service Based Architecture (SBA) in the network core,[14]

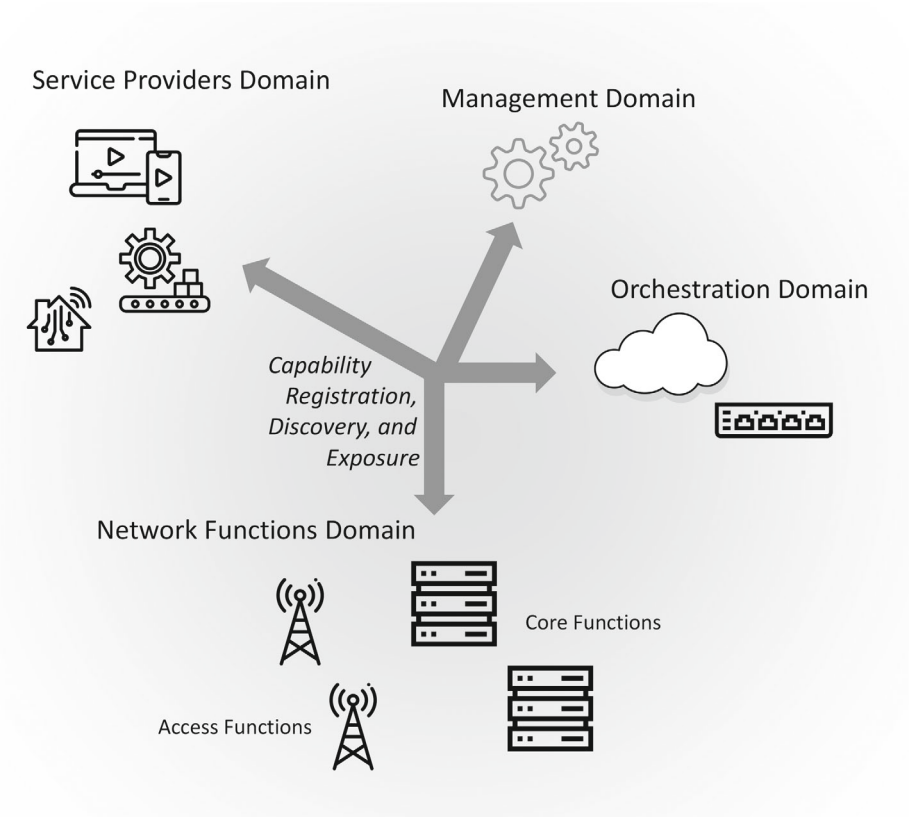

FIGURE 1 The domains of a 5G network

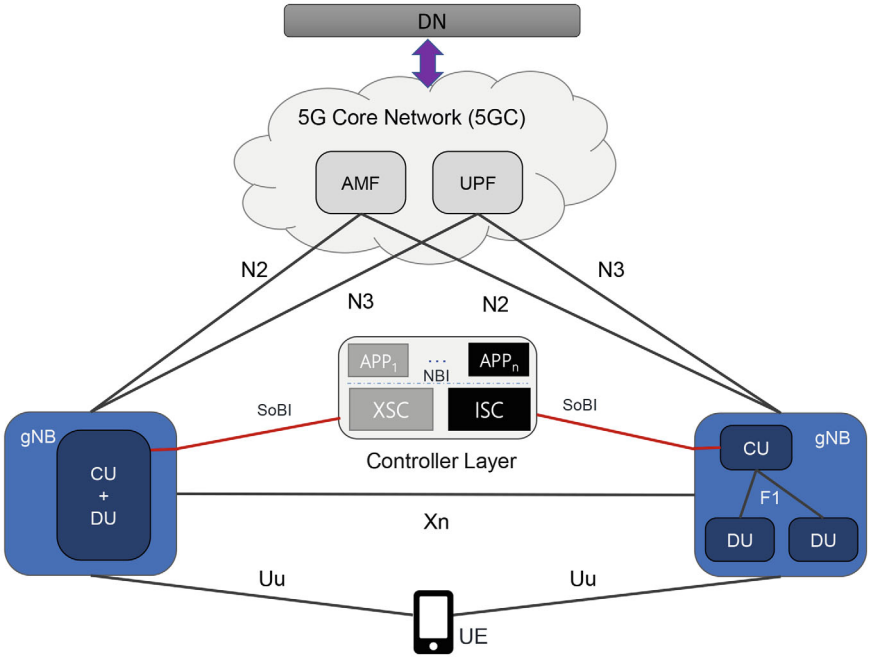

FIGURE 2 Softwarization of the RAN controller, as proposed in Reference 20

and the *xApps* defined in the O-RAN controller hierarchy.[15] However, the flexible interaction guaranteed by the open interfaces devised by these approaches are traditionally confined into one specific domain (eg, network functions), while the software landscape in the context of mobile networks is much broader, as discussed next (Figure 2).

2.1.1 | The network function domain

The main functions and interactions of the functions within the RAN and between RAN and CN have already been specified by 3GPP Release 15.[16,17] 5G new radio (NR) includes the service data adaptation protocol (SDAP) layer in the user plane, which enables the mapping of QoS flows to the radio bearers increasing the degree of freedom for QoS enforcement in RAN, and the F1 interface which enables the central unit-distributed unit (CU-DU) split between packet data convergence protocol (PDCP) and radio link control (RLC). The CU-DU split is a step toward the flexible centralization of RAN functions and network function virtualization. The 5G base stations, called gNBs, are interconnected via the Xn interface, which is used, for example, for handovers. N2 and N3 are the interfaces toward the 5G core network (5GC), which define control plane and user plane procedures, respectively. Further, fundamental support for network slicing is provided in the RAN, where slicing awareness in RAN is attained via Network Slice Selection Assistance Information (NSSAI)[18]), including one or more Single NSSAIs (S-NSSAIs), which allow to uniquely identify network slices and hence the associated configurations.[16] Leveraging on this architecture, the O-RAN Alliance promoted a much more flexible layout based on the RAN Intelligent Controller (RIC), an element that abstracts the complexity of the distributed and centralized units, allowing operators to implement and operate custom control plane functions.

Since 3GPP Release 15, 5G CN (5GC) follows new design paradigms with control plane–user plane (CP/UP) separation, modular NFs, and SBA. The 5GC NFs and network entities (NEs) including CP functions (eg, AMF, SMF, PCF, AF, UDR, etc.) and UP functions (eg, UPF) are specified in Reference 3. In addition to the NFs/NEs used for conventional mobile network control (ie, UE identity, access, mobility, session, and policy management, etc.), the 5GC comprises also new NF/NEs supporting service-based communication between 5GC CP NFs/NEs, such as the network repository function (NRF) and network exposure function (NEF), that are specified to enable the service registration and service discovery of the 5GC CP NFs/NEs in the same domain (eg, 5GC of a network provider).

The providing NF/NE registers its services as well as the addresses to reach these services in the NRF using the service registration procedure.[19] Then, the consumer NFs/NEs are able to discover the services provided by other NFs/NEs by an inquiry to the NRF using the service discovery procedure.[19] Alternatively, the management domain can also configure directly in the NRF the services provided by a certain NF/NE as well as the address to reach them.

In some cases, the 5GC CP NFs/NEs may need to communicate with some other network domains (eg, application functions (AFs) provided by a 3rd party, applications, V2X servers, or a multiaccess edge computing (MEC) platform). For such cross-domain communications, special service communication restrictions and information translation are necessary. 3GPP Release 16 specifies in Reference 3 that NF capabilities and events may be securely exposed by NEF for example, 3rd party AFs and MEC. In the other direction, NEF is responsible for the secure provision of information from external applications to the 3GPP network. An example of this includes data collection from an external AF for network data analytics,[14] UE parameter/service parameter provision.[19] Besides the cross-domain security guarding tasks, NEF performs also the translation of internal-external information in different domains (eg, slice identifier, individual/group identifier, address, location information, etc.).

Network Data Analytics Function (NWDAF) is another new NF introduced in 5GC. NWDAF represents operator managed network analytics logical function. In 3GPP Release 15, NWDAF provides network analytics on the load level of a NF. In 5GC SBA, data analytics services of NWDAF can be consumed by any NF. 3GPP Release 16 extends the usage of NWDAF also to use cases beyond load level, for example, network performance analytics, slice load level related network data analytics, observed service experience related network data analytics, UE related analytics, quality of service (QoS) sustainability analytics, etc. A more detailed NWDAF framework has been specified, covering the data collection and analytics provision from/to NFs in the same network, from external AF, and from OAM.[14] NEF can expose the analytics generated by NWDAF to AF, and NWDAF could also collect data from OAM.

2.1.2 | The management domain

Starting from 3GPP Release 15, the 3GPP operations, administration, and maintenance (OAM) domain, in the following also referred to as management plane, has introduced the Service-Based Management Architecture (SBMA). In this framework, a management service offers management capabilities. The most essential management services include generic provisioning, fault supervision, and performance assurance management services, which are typically produced by the NF or a lower management layer (eg, Network Function Management Function or Network Slice Subnet Management Function). These management services are accessed by management service consumers via a standardized service interface composed of management service components that describe the management operation, the managed entity, and the managed data related to this entity, for example, performance information. Generally, management services can be exposed to any authorized consumer. To allow for (eg, policy-based) access restrictions, 3GPP SA5 has introduced the Exposure Governance Management Function (EGMF), a management function responsible, for example, abstraction, simplification, filtering, aggregation and so on of management services, incl. data services.

Additional management service abstraction may be needed because of a lacking trust relationship between management service producer and consumer, for example, if the service consumer resides outside the operator's administrative domain. The business model "Network Slice as a Service" (NSaaS), as defined in Reference 21, requires specific sets of management services to be exposed to (and consumed by) an operator's customer. 3GPP[21] also lists lifecycle management services for a communication service instance, for example, activation, modification, management data analytics (MDA)-assisted service-level specification (SLS) assurance, termination. The MDA service (MDAS) hosts the capability to process and analyze (raw) network data (eg, performance measurements, trace reports, QoE reports, alarms, configuration data, other network analytical data, etc.) to detect specific events, make predictions about network performance, etc. The provided analytics reports may include recommended actions, as well.[22] Such MDAS instances cannot only be tailored for a specific use case, but they can also be exposed, via EGMF, to external consumers that may subscribe to customized analytics reports.

2.1.3 | The orchestration domain

While the traditional network domains (ie, NFs and management plane) were already present in legacy networks (ie, up to 4G/LTE), their Orchestration and the (advanced) configuration experienced an incredible boost, mainly due to the introduction of the SDN, NFV, and containerization technologies:

- *Resource virtualization* allowed for on-demand provisioning, following the very successful paradigm of Infrastructure as a Service (IaaS), offering thus the seamless opportunity of creating, re-configuring, and terminating network services.
- The success of *network programmability* boosted the adoption of API-based access to network configuration, enabling a flexible (re)-configuration of NFs, in contrast to traditional network control (eg, manual CLI-based approaches).

Given its scope of software-driven, application-agnostic management of general-purpose cloud resources, orchestration and lifecycle management procedures have not been tackled by 3GPP. For instance, the orchestration of VNFs or containers is not being included in the 3GPP work. As a matter of fact, orchestration of network resources is currently achieved through vendor-specific, standard-compliant solutions such as Nokia CloudBand Suite* or open source initiatives either hosted by standardization bodies (such as the implementation of ETSI NFV MANO[23] provided by OSM[24]) or other fora (eg, ONAP†). While the aforementioned solutions deal with the orchestration of cloud resources, similar solutions for radio resources are less widespread. For example, industry initiatives such as O-RAN have defined radio orchestration procedures.

Besides orchestration, these resources also need to be configured according to network conditions, like load and traffic patterns. A relevant example is the configuration of the inter- and intra-datacenter networks (eg, through an SDN controller) or the transport network[25] (eg, by using an SD-WAN approach). Such a controller-based approach, which is followed by SDOs, can also be extended to the network domain, as also proposed by O-RAN for the access network.

2.1.4 | The service provider domain

The novel network softwarization paradigm introduced by 5G and beyond 5G networks allowed for a diverse and heterogeneous landscape of tenants.[27] That is, different service providers such as industrial verticals, are now a fundamental piece involved in the network operation, in clear contrast with what has been traditionally happening up to the legacy 4G networks.

The 4G network provisioning is characterized by a full over the top (OTT) service delivery model. Instead, the new model can provide a more integrated view of the system from the tenant and service provider perspective, who can, by leveraging on novel configuration primitives, act on the underlying network slices. From the standardization point of view, this concept has been increasingly attracting more attention. For instance, 3GPP[21] currently defines two management models: network operator internals (ie, the legacy solution, in which the tenant has no visibility on the underlying system) and NSaaS in which the tenant can manage the network slice as manager via an exposed management interface, and optionally provide network slicing capabilities to other tenants.

Furthermore, industrial initiatives such as 5G-ACIA[26] describe the requirements for a 5G exposure reference point toward enterprise tenants to promote better integration between 3rd party service providers and network operators. Envisioned for Industrial Internet of Things (IIoT) use cases, the concept currently under development in 5G-ACIA, depicted in Figure 3, proposes the exposure of selected functionality from a 5G nonpublic network or from the UE toward IIoT applications in the IT enterprise domain. Among these exposed functionalities we have

- The exposure of selected 5GC control plane functionality (eg, NWDAF data analytics)
- The exposure of the 5GC capability as a whole (eg, provide a specific QoS for an AS session, influence traffic routing, provision service/UE specific information, etc.)
- The exposure of the 5GS management capability (eg, slice configuration)

However, 5G-ACIA does not mandate any specific solutions to be used for the exposure reference point. Rather, usability, simplicity, modularity, and extensibility are listed as key requirements. The specific capabilities to be exposed

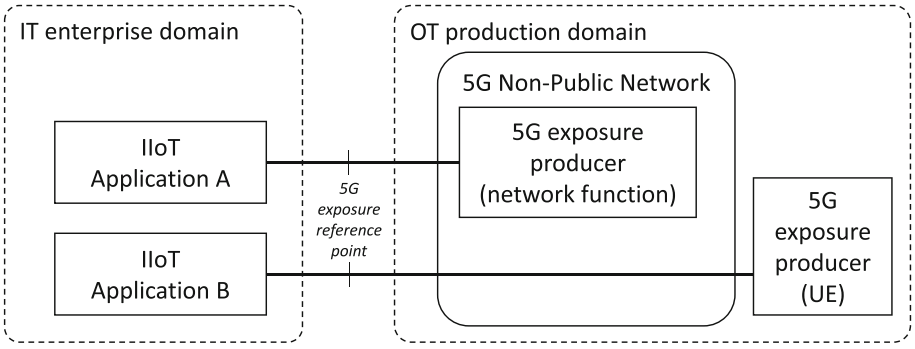

FIGURE 3 5G-ACIA[26] concept for 5G exposure reference point

depend on the envisioned use case. Among them, device provisioning and onboarding, device connectivity management, device connectivity monitoring, device group management, device location information, network monitoring, and network maintenance can be listed. Thus, the proposed exposure framework shall also take into account the requirements of 5G-ACIA. Currently, 3GPP only envisions the traditional network operator model with minimum interaction of NFs with external functions and the NSaaS model. As for example, 5G-ACIA requirements show, more flexible interfacing may be beneficial for different business models (eg, hybrid private public deployments).

Finally, we remark that the quest for a tighter interaction between service providers and network operators is also pursued by industrial consortia besides beyond the factory ones: the *NetApps*[28] concept introduced by many research projects nowadays is also promoting the usage of open APIs between these traditionally separated domains. Among them, we can certainly consider new providers that make use of edge computing technologies, for example, offloading use cases that require very low latency, that is the case also for other providers such as vehicular ones, to provide example autonomous driving services. In this case, the open interactions between such applications and the underlying access technology[29] is paramount. For instance, as already discussed in Reference 30, the integration of MEC with relevant architectures such as O-RAN is still being discussed, and still, these efforts are targeting a specific architectural element (the O-RAN xApp controller, but it lacks exposure functionality toward other domains, such as the Core.

2.2 | The capability types

As previously discussed, although different elements of the state-of-the-art network architectures already provide data-driven functionality, their limited scope (very often bounded to one specific domain) may hinder the automated operation of the network that involve cross-domain activities such as reactive orchestration upon network triggers, or service-driven network re-configurations that are at the basis of the autonomous operation of the network.

Thus, owing to the data-driven vision of the network, we introduce four types of capabilities that summarize what kind of interaction shall happen across domain boundaries. Cross-domain interaction shall be enabled by inter-domain message buses. Please note that registered capabilities are associated with authorization levels for potential consumers from the same and other domains.

Capability type 1: Monitoring and data collection: This capability is related to the provisioning of raw monitoring data from and to different Network Elements, and it is already implemented for some network domains. For instance, NFs can provide KPIs related to the cell performance (eg, handover failure rates, cell load, etc.), user-centric KPIs (eg, per user throughput, latency, etc.) and KPIs related to the end-to-end service performance. Also, monitoring data of the infrastructure (eg, CPU and RAM utilization) falls into this category. Finally, this capability type includes the capability to define customized measurement jobs, trace collection configurations, or real-time performance measurements on the monitored element.

Capability type 2: Triggers, alarms, and fault supervision: While the data collection capability discussed before provides a way for collecting information from Network Elements at full granularity, there is the need for a more refined way of accessing it. That is, most of the elements in the network need to react according to well-defined state machines

upon triggering events. This includes the normal network operation (eg, RRC state handling and radio link failures), but it is of particular importance when dealing with fault supervision. Thus, providing event notifications, such as, alarm information, alarm state changes, alarm correlation information can be categorized in this capability type, which can be further enriched to include notifications on associated trouble-shooting actions.

Capability type 3: Actions, control, and configurations: Besides exposing data as well as alarms and events, Network Elements of different domains shall also expose configuration and control capabilities to other elements. Generically, this capability type comprises capabilities to act, that is, to create, modify, delete objects as well as their parameters and configuration attributes. The definition of an object then depends on the specific domain that is exposing this capability. For instance, in the orchestration domain an object is, for example, a VNF, a resource, a network service and so on. Similarly, in the network domains, an object is, for example, a parameter of an NF or a network slice instance.

Capability type 4: Network intelligence and policy recommendations: Future network operations will be intelligent. This means that most of the tasks that currently require human intervention to achieve optimality in the network will be handled automatically and with some kind of AI in the loop. Thus, Network Elements shall expose the capability of performing complex analytics on inputs coming from other elements (ie, the ones exposed by capability types 1-3 above). For instance, intelligent elements in the networks shall include root cause analysis or impact analysis. Individual data analytics capabilities can be combined to process incoming events and other information related to faults and performance of the monitored objects, aggregate, and analyze the information and derive novel information required for enhanced operations, for example, failure prediction to prevent faulty network states. Policy recommendations exposed through this capability are eventually consumed by other elements in each domain or across domains

2.3 | Network data exposure beyond standardization

With the increasing interest on network automation research topics, also the research in the field of network data exposure has grown. However, this topic has been mostly addressed from a very specific point of view: network security.[31]

Indeed, the security aspect is paramount when opening up information between network domains that have been traditionally separated. Moreover, this problem is made even more complex by the fact that those domains may be operated by different operators and tenants.[32]

These works are however complementary to ours, as the interaction between consumers and Registry and Exposure functions can be made secure with state of the art solutions for authentication and encryption.

A work similar to ours is the one in Reference 31 which, however, focuses on just the core domain, enhancing the current NWDAF-based solution. Still, one of the most important areas of interest for network exposure research is the one targeting service providers, as discussed in Reference 33. As we also discuss in our experiment, exposure functionality for autonomous networking is very beneficial for private deployment of 5G networks, where service providers can create control loops between application and network deployments.

3 | TOWARD A WIDER EXPOSURE FUNCTIONALITY

By opening a flexible exposure of capabilities among domains, as discussed in Section 2, mobile network systems can implement closed-loop control system, following a producer-consumer or publish-subscribe approach. Following the recent efforts in providing network automation for the major operational tasks such as for example, service and resource orchestration, and artificial intelligence is deemed as one of the most important tools to achieve this vision. In particular, machine learning (and especially deep learning) solutions are gaining a lot of attention in both industry and academia in the last few years. Thus, to enable network automation through the usage of such solutions, very large amount of data is needed. That is, the correct training and configuration of these models require the availability of diverse and rich data, which is likely to be produced by network functions running in one of the other domains, as discussed in Section 2. Hence, in this Section, we summarize the most important network data exchanges that happen among domains. These interactions are summarized in Table 1 and deeply discussed next, by producing domain.

TABLE 1 Producing domains and their capabilities

Capabilities/Producingv network domains	(1) Monitoring and data collection	(2) Triggers, alarms, and fault supervision	(3) Actions, control, and configurations	(4) Network intelligence and policy recommendations
(A) Network functions	NW resource utilization	NW resource failure	NRM parameters	NWDAF
	UE traffic conditions	QoS unfulfillment	Procedure (ICIC) parameters	RAN Analytics
	UE counters	Network functions SW exceptions	Mobility management	Long term RRM
	5GC counters			
(B) Management	Cell Traces	Cell outages	SON	MDAS
	Network slice counters	Slice-level SLA failures	lice lifecycle management	
(C) Orchestration	NFVI monitoring	NFVI alarms	VNF placement decisions	AI as a service
	WAN monitoring	WAN links failures	VNF deployment flavor	VNF placement algorithms
				Root Cause Analysis
(D) Service providers	Manufacturing process monitoring	Manufacturing line failures	Production cell layout reconfigurations	Service domain analytics
	Application service status	Massive churn rates	Expected traffic patterns	Business intelligence

Network functions: In the context of mobile network management and orchestration, the collection of raw counters and aggregated performance metrics from NFs in both RAN and CN domains is a feature that is currently supported by many functions (cell A1 in Table 1). In legacy networks, this task is typically performed through the point-to-point interfaces between an NF and its Element Manager (EM) and, if applicable, its Virtualized Network Function Manager (VNFM). Continuous monitoring data coming from network probes belong to this category, however such data gathering points are often vendor-based as 3GPP does not define the specific interfaces between an NF and its EM. Moreover, this data is usually not fine grained both for the kind of data (ie, counters only) and temporal resolution for example, several minutes and most importantly it is not delivered in real-time, typically through file transfer. This is also captured by some Open Source implementation such as Linux Foundation magma[‡], which exposes these metrics toward the orchestration.

Similarly (cell A2 in Table 1), some standardization effort such as the one carried out by GSMA, has defined the Generic Slice Template (GST)[34] which provides the attributes, including KPIs (eg, maximum number of UEs, maximum number of PDU sessions, Maximum aggregated UL/DL bit rates, etc.) a network slice should fulfill. When a GST is filled with values, that is, with the customer requirements, the network slice type (NEST) is constructed. 3GPP specifies that network slice management function (NSMF) is responsible for the network slice commissioning following the service level agreement (SLA) between the operator and the customer. Considering the SLAs, the NSMF can expose policies to the CP for SLA monitoring and SLA fulfillment. The CP NFs monitor whether the SLA quotas are exceeded or not and expose that information to the management layer. At the same time, the CP NFs may enforce the quotas by means of rejecting registration/session establishment requests exceeding the quota.[35] All these events fall into the alarms category.

Software exceptions (eg, issues with the code) are another category of alarms. While they have already existed in early networks with proprietary, fully integrated software and hardware components, they have significantly proliferated and become more inter-operable in the era of cloud networking. This alarm type is usually consumed by management functions that can decide to for example, downgrade the software and rollback to a more stable release.

Control plane NFs and exposure of their capabilities (cell A3) have received great attention in mobile network standardization. 5GC already provides a mechanism for dynamically accessing CP services for NFs through the Network Exposure Function (NEF) to the applications. Among the parameters that can be configured, we can list the UE-related

ones, through the UDM, but in general, any parameter that is included in the information model specified by SA5 in Reference 36. Access functions also expose configuration capabilities (used by management functions mostly), such as the parameters currently used by SON functions like eICIC, mobility, or load balancing. Finally, specific configurations can be exposed directly to the service provider, as currently envisioned by 5G-ACIA or by 3GPP SA6 with the work on mission-critical applications.

Also starting from 3GPP Release 15, the mobile network started to provide analytics (cell A4) services, mostly through the NWDAF in the 5GC, which, for example, provides information about the load of a NFs which may be used by other NFs to adjust sensible settings for that is, load balancing purposes. Similar data analytics elements are not available in the current 3GPP RAN architecture, but are envisioned by other initiatives, such as O-RAN through the Radio Network Information Base (RNIB).Those data analytics services are also consumed by management functionality for network slice related decision.

Management: Management functions (cells B1-B2), instead, can expose monitoring data at any of the granularity levels that is used for handling the lifecycle management of objects (service, network slice, network slice subnet, and NF), providing counters regarding for example, UE-related events, cell loads, or network slice load for management and orchestration purposes.

Traditionally (cell B3), the interaction between the service providers and the network management happens through the OSS and BSS, which are usually involving customer care services and require "human-alike" timings. Instead, by offering the service provider the capability of directly influencing the lifecycle management, the duration of such a process can be reduced very much and allow timely network management upon changes, without indirect policy configurations.

Finally (cell B4), the management system also produces analytics information through the Management Data Analytics Service (MDAS). This service is consumed by other management entities or forwarded to other domains such as the orchestration or the network function domain. Still, in current standardization efforts, this interaction is limited to very few metrics.

Orchestration: State of the art SDN and NFV orchestration technologies such as Open Source MANO (OSM) or selected ONAP components, already provide a very high number of continuous statistics (cell C1), mostly provided by the Virtual Infrastructure Manager (VIM) elements. Metrics such as utilization of CPU, RAM, and networking resources are usually provided as time series data. Such kinds of orchestration systems can also trigger alarms (cell C2) when malfunctions happen (eg, nodes running out of memory) or the underlying network infrastructure (eg, generic hardware faults in switches, routers, and gateways) detect a failure in wide area links.

Also, orchestration systems provide rich APIs (cell C3) that are leveraged to perform network (slices) lifecycle management, allowing operations such as VNF re-orchestration and re-configuration, and also path management of the physical and logical links that interconnect different infrastructure deployments located far away. Finally, orchestration and control solutions usually provide (cell C4) "AI as a Service" features (for instance, the inclusion of ACUMOS[§] into ONAP) or analytics services such as root cause analysis or VNF placement suggestions. Summarizing, while there is a large amount of data generated and consumed by the orchestration domain, most of it is restricted to the domain. Very recently, Linux Foundation has also launched the Akraino[¶] project, which aims at simplifying the interaction among service providers and edge orchestrators, through the usage of blueprints.

Service providers: The interaction between the (application) service provider or a stakeholder from vertical industry and the network is currently very limited, as the traditional business model, with over-the-top (OTT) services, was not directly considering this option. However, a service provider can provide countless capabilities to other domains. Clearly, real-time metrics (cell D1) from the applications such as the monitoring of manufacturing processes, the audio-video production quality metrics, or generic service status. Similar considerations apply to alarm (cell D2) capabilities like assembly line failures. failure of live coverage or notifications of massive user churn rates at the application layer and increased delay.

The main problem to solve in this context is related to semantics, as applications can be disparate, and the related metrics can be difficult to categorize. An ongoing effort in this sense is being carried out by 3GPP SA6 with the specific study items on drone communications[37] and support for "factory of the future" applications.[38] On the other hand, the service providers can offer re-configuration (cell D3) capabilities for some of their applications (such as reconfiguration of the production layout in an industrial environment), or analytics and intelligence services (cell D4) to the infrastructure network operator provider which are, again, dependent to the kind of running application.

In a similar way, the network domain(s) can expose a higher number of capabilities to the service provider domain. This applies to both monitoring and KPI collection capabilities as well as selected NF configuration and control possibilities allowing the service provider to adapt the network resources within predefined boundaries.

4 | THE ARCHITECTURE

In order to support the data-driven and open design of the network discussed in Section 3, we propose extensions to the state-of-the-art 5G architecture that fully embody the capability exposure features previously discussed. The proposed unified service-based architecture relies on three types of components that provide the needed functionalities: the inter-domain message bus, the capability registration and discovery functions, and the enhanced exposure functions. The overall architecture is depicted in Figure 4, which joins together the different domains discussed in Section 2 into a common network exposure framework. As discussed next, our proposed framework leverages on an enhanced capability registration and exposure functionality that shall be available at each domain. Specifically, in Figure 4, we depict with solid black boxes the elements that are already envisioned by the SDOs in each domain, while red boxes and blue boxes represent the discovery/registration and exposure functions, respectively. Some elements are already integrated in the SDOs architectures (such as the cases of the 5G Core Network functions, while others shall be integrated in the SDOs proposals. In the following, we introduce them and discuss their integration into the overall architecture of a 5G and beyond) system.

4.1 | Capability registration, discovery, and exposure through the inter-domain bus

Enabling closed control loops among different domains means enabling a flexible way of exchanging capabilities among them. This goes beyond the current standard architectures, which often rely on a standalone design in each individual domain, not fully capitalizing on possible advantages discussed in Section 3. Our proposed architecture thus adapts, extends, and integrates current service-based approaches for different domains (ie, the 5GC, the 5G Management System, and ETSI NFV MANO) by means of an inter-domain message bus that connects them, enabling the capability exposure and consumption across domains. All the domains that use capability exposure interface the inter-domain bus with two types of functions, that is, they shall be included in each of them: the capability registration and discovery function and the exposure function:

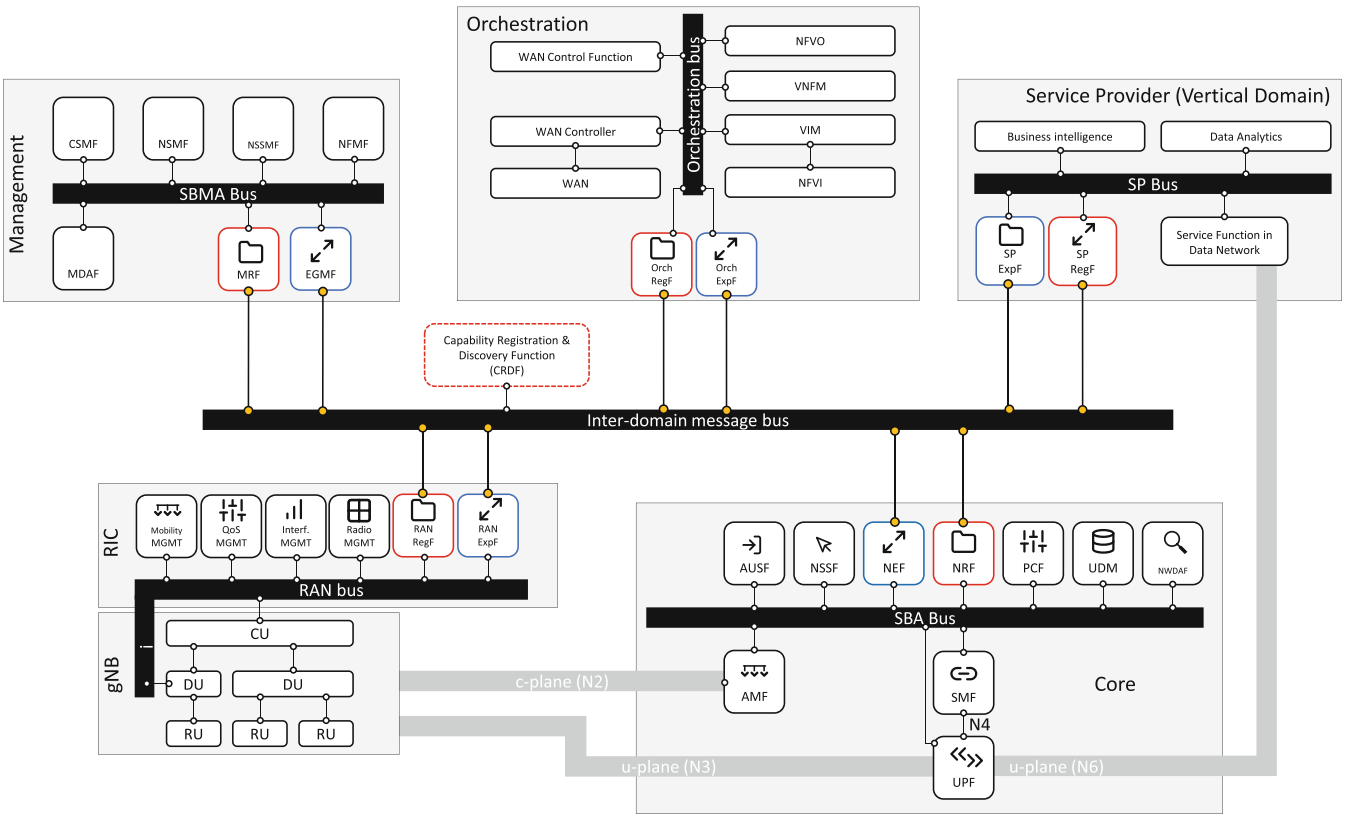

FIGURE 4 The proposed unified architecture for cross-domain capability registration, discovery, and exposure

The capability registration and discovery function (CRDF) provides a registration service for available capabilities within each domain and, if required, also capabilities exposed by other domains using the inter-domain bus. It maintains a list of registered capabilities and responds to the service discovery inquiries from the consumers in its domain. The CRDF also acts as a proxy for discovery requests, when these are targeting capabilities that are available outside the source domain. Capabilities can be listed using a taxonomy such as the one we propose in this article, following an approach similar to the one used by NWDAF to list the available analytics within the 5G Core domain, and implemented using frameworks such as the OpenAPI[#] one.

The exposure function makes a capability available to other consumers both within the same domain (as it is partially the case with state-of-the-art technologies) and to other domains. The exposure function can enable access-controlled communication across different domains.

The aforementioned functionality has to be provided within all the domains. In the 3GPP CN domain, NRF and NEF already provide similar features within the CN and between the CN domain and Service Provider domain (for instance, to allow data collection from application to the NWDAF or analytics exposure from NWDAF to the application), so they have to be extended to support the access by other capability consuming functions (CCF) from other domains and allow the exposure toward them. For the radio access network domain, O-RAN architecture already realizes selected features of the proposed framework. However, O-RAN does not provide exposure, registration, and discovery functions beyond the interfaces toward the orchestration domain. Additionally, other open source implementations of the RAN network functions, such as OpenAirInterface (OAI) are also not providing this kind of functionality.[39] Thus, the exposure, registration, and discovery functions can be implemented through specific applications from the non-real-time controller that then interfaces with the other (both real-time and nonreal-time) applications. The RAN bus can also interface with the (virtualized) access domain functions, for example, the centralized unit (CU) of the 3GPP architecture, compare Figure 4.

On the network management side, an existing module in the 3GPP architecture (ie, the Exposure Governance Management Function, EGMF) can be extended in a similar way as the NEF (ie, to support the exposure of capabilities to other domains). Nevertheless, an additional registration module needs to be included, as it is currently not (yet) specified by 3GPP SA5 in the required manner. Then, these elements interact with other management entities through the Service-Based Management Architecture (SBMA) bus.

The orchestration domain, as specified by standard developing organization such as ETSI NFV, currently does not provide an exposure and registration functionality. Hence, such functionality needs to be added entirely to this domain. Also, as the ETSI NFV architecture currently uses peer-to-peer reference points, we envision an orchestration bus to provide capability consumption within the domain. This includes elements such as NFVO, VNFM, and VIM, but also SDN controllers and WAN controllers.

Finally, in the service provider/vertical domain, these elements will need to be added as well to support the proposed architecture framework vision. This could enable, for example, analytics exposure between the service provider domain and the network domains. As discussed before, an additional capability registration and discovery function, operating at intra-domain level may be added. This allows registering functions of each domain making them mutually discoverable. Alternatively, registration and discovery functions of each domain can discover themselves dynamically through the inter-domain bus.

The most similar concept compared to ours is the one promoted by ETSI ZSM,[40] through the *integration fabric* across different orchestrators. However, there are two fundamental differences between our proposal and the work done by ETSI ZSM: (i) the ZSM proposal mostly focuses on the synchronization between management, orchestration, and network control, while we also include user plane functions and, especially, service providers. As discussed previously, the openness of the system toward service providers is a fundamental feature for next generation networks. Then, (ii) ETSI ZSM proposal is highly hierarchical, with the management domain playing a master role in the system: our framework instead fosters an open exchange among domains, leveraging network programmability concepts, without a predominant role of the party that operates the management domain.

4.2 | Procedures

In the following, we briefly discuss how the most fundamental procedures can be achieved using the proposed architecture.

Capabilities (de)registration: The registration functions that operate in each domain take care of maintaining a current list of consumable capabilities. Thus, the APIs register, deregister, update, and notify, shall be extended with the capabilities that have to be registered. These APIs shall be available also on the inter-domain bus.

Capability discovery: Every time an NF (at each domain level) needs to discover a capability from another function, it shall follow a two-step procedure:

1. It requests the capability at its intra-domain CRDF, using a discovery API for the needed capability.
2. If the capability can be served intra-domain, then the CRDF points to the capability producing function (CPF). There, the requesting NF can request the needed capability, for example, by means of subscription procedures.
3. If the needed capability is produced outside the domain, then the registration function contacts the inter-domain CRDF or CRDF in other domains and obtains information about the external exposure function.
4. The capability exposure is finally handled by the external exposure function.

Capability exposure: The capability exposure from each domain is performed through the exposure function that runs therein. The module receives a request from another function, either directly (if the exposure has to be performed within the domain) or through another proxying exposure function (in case of an inter-domain request). Here, the subscription request is then forwarded to the producing function, which finally provides the capability.

Routing, forwarding, and load balancing: In order to support the mechanisms discussed previously, the CRDFs and the exposure functions perform request routing and forwarding of capability discovery or consumption requests across the domains. Then, the final capability exposure will happen directly between the different entities: if end-to-end connectivity is available, then the communication is direct, otherwise, the registration functions of the two domains take care of routing this traffic as well. Finally, CRDFs and exposure functions of each domain handle load balancing by pointing the requests to the final instance of the NFs following for example, a round-robin policy.

Capability consumption: Communications between CPFs and CCFs happen by means of one of the following two modes:

- *Request/response*: This is relevant especially for capability Type 3, in which NFs expose their configuration parameters to external consumers. This can be achieved by, for example, HTTP(S) requests.
- *Publish/subscribe*: Other capability types can be consumed following this paradigm using message-oriented implementation such as Apache Kafka or Rabbit MQ.

The consumption procedure may happen in two ways: the exposure function can directly hand off the capability consumption to the consumer function or handle it directly. In the latter case, as the complexity is handled by the exposure function, no protocol translation is required among domains (which use the exposure function as proxy).

Access control: Authentication and authorization of CCFs: Capability discovery and consumption requests are handled in the first place by the CRDF, which keeps the authorization information regarding CCFs for each registered capability. Such capability authorization information is configured as one of the components of the registration profile of capability producer in CRDF. The profile should include the CCF type(s) and CCF realms (the domain of origin) that are allowed to consume the registered capability. Moreover, CRDF also performs CCF authentication in order to verify the identity of the requesting entity, for example, using Extensible Authentication Protocol (EAP). This latter point is of particular importance when different service providers want to gain access to the capability exposure (which may be part of a subscription plan offered by the network operator).

5 | USE CASES

In the following, we describe three use cases that leverage the architecture presented in Section 4, describing how they can use the proposed architectural procedures. We generalize in Figure 5 the roles of the capability producing function (CPF) and capability consuming function (CCF) for the sake of clarity.

5.1 | Edge applications for service provider integration

This use case describes a scenario where an edge application such as one of the proposed by the ETSI multiaccess edge computing (MEC)[41] shall interact with a 3GPP network. Here, cell trace data are produced in the RAN

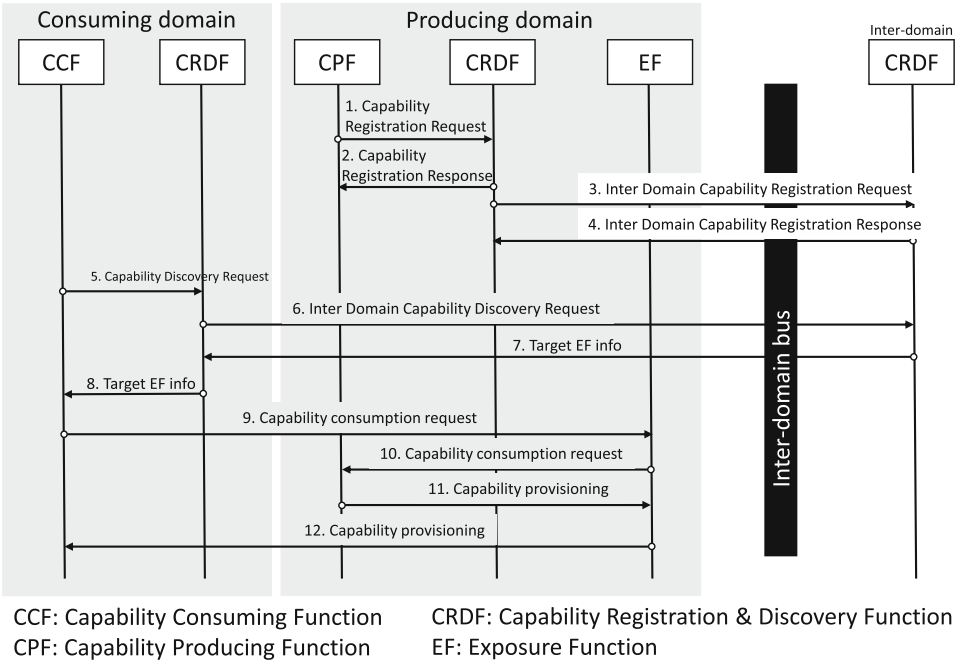

FIGURE 5 The network capability exposure and consumption from/by different network domains

domain and first consumed by a management domain function. There, data is further enhanced, for example, by means of data analytics, which allows for offering an according management domain capability. This can then be consumed by entities outside the 3GPP operator domain, for example, an Edge Application in the service provider/vertical domain. To this end, we take the trace management as an example to illustrate a functional interaction.

In order to realize this scenario, the RAN cell trace provisioning capability needs to be registered in a first step. Second, the network management domain function, that is, enhanced trace collection entity (eTCE), needs to discover the according capability before consumption. Third, the management capability of enhanced data analytics needs to be discovered by an external SP domain entity, that is, by edge application. These three basic steps are outlined in more detail in the following.

1 *Registration of RAN capability*:
- (a) The CPF, that is, gNB-CU registers capability (of type 1) "producer of raw trace data for cells A, B, C", incl. metadata of capability, at RAN-domain CRDF. This corresponds to steps 1 and 2 in Figure 5.
- (b) RAN-domain CRDF decides to register capability in inter-domain CRDF, incl. authorization information. This corresponds to steps 3 and 4 in Figure 5.

2 *Discovery of RAN capability by the management domain function*:
- (a) CCF, that is, eTCE, issues discovery request for capability "raw trace data for cell A, B, C" to management-domain CRDF (step 5).
- (b) Since no such capability is registered at the domain-level CRDF, CRDF forwards discovery request to the inter-domain CRDF (step 6)
- (d) Inter-domain CRDF authorizes eTCE according to information provided by RAN-domain CRDF and provides a Unique Resource Identifier (URI) of target EF as well as a security token to management-domain CRDF (step 7)
- (d) The management domain CRDF forwards target EF information and security token to eTCE (step 8)
- (e) eTCE sends capability consumption request (here: subscription request for trace data) to target EF (incl. security token and delivery URI of eTCE) (step 9)
- (f) Target EF (ie, RAN-domain EF) authorizes request and forwards it to CPF (ie, gNB-CU)
- (g) gNB-CU provisions capability via RAN-domain EF to CCF (ie, eTCE) (steps 11 and 12)

3 *Discovery of management domain data analytics capability by edge application in SP domain*:
(a) CCF, that is, the Edge Application discussed here requests producer of capability "trace data analytics for cell B" at SP-domain CRDF (step 5)
(b) The Service Provider domain CRDF forwards capability discovery request to inter-domain CRDF (step 6)
(d) The inter-domain CRDF forwards discovery request to CRDFs in other domains, including management domain
(d) Management-domain CRDF checks if requesting consumer is authorized for capability consumption; if yes, a security token is generated
(e) Management-domain CRDF provides target EF information and security token to requesting CRDF; these are further forwarded CCF, that is, to EA (steps 7 and 8)
(f) EA sends capability consumption request and security token to management-domain EF, which forwards it to CPF. CPF provisions capability to EA via EF (steps 9-12)

5.2 | An enriched interface for service providers

The integration between the service providers and the network operators envisioned by 5G networks is yet to be completed. services are currently provided over-the-top (OTT). However, next generation mobile networks will transition to a fully blended approach, in which service provider and network provider cooperate to fulfill the requirement of the service. Hence, the interactions between them shall be provided with a flexible and customizable interface, leveraging on an enhanced capability exposure, for all the types defined in Section 2. In order to introduce successful business models such as the one currently employed by Amazon AWS, Google Cloud Platform, or Microsoft Azure, where the user can freely configure the characteristics of the service they are running, additional capability exposure are needed:

- Amount and kind of resources: in order to define details such as the number of nodes running a network service, their power, if they need or not special hardware (such as a GPU, a FPGA, or a Smart NIC), the service provider needs to gain access (through secure and validated APIs) to the Orchestration platform. This can happen directly or through a more high-level (eg, intent-based) interface.
- Continuous flow of metrics from the network, including standard QoS metrics (such as latency, aggregated bandwidth, number of users) but also the capability of gathering user defined metrics, by merging more simple metrics at once.
- Receiving suggestions from the network operator regarding business strategies, including information about subscription costs and dynamic SLA (eg, quality upgrades) or proposal for re-orchestration of NFs (scaling up or down, relocation to the edge).
- Link raw data analytics from the network infrastructure (including predictions) with the internal business logic of the service provider, which can then make decisions about how to reshape the service offer.

In the following, we exemplify this with a video streaming service provider using a richer interface that introduces several advantages compared to OTT operation, (following the steps depicted in Figure 5).

In this case, three domains (the network functions, the management, and the orchestration) are still under the same stakeholder (ie, the network operator) which offers access to the inter-domain bus (through, eg, an IP network) to the vertical service provider.

Thus, the service provider can subscribe to continuous monitoring (capability type 1, steps 5-12 in Figure 5) from the Orchestration and Network Functions domain to observe the current KPI and map them with QoE metrics available at the application layer.

The service provider also subscribes to dynamic SLA events (capability type 2) coming from the operator BSS Management, which informs about different subscription options. For instance, the Video Service Provider could be offering the service using a "Silver" plan subscription, and the Operator could offer an upgrade to a "Gold" plan (with the information of the increased cost and the QoS improvements) or the downgrade to a cheaper and less powerful plan.

Finally, through re-orchestration suggestions (Capability type 4), the service provider has access to a rich version of the traditional OTT paradigm, and also go beyond the NSaaS paradigm. So, the service provider (by accessing the configuration parameters, offered through the capability type 2, steps 1-4 in Figure 5) can apply its business logic, taking

re-orchestration and dynamic SLA decisions based also on metrics such as the current revenue flows, the subscribers' churn rate or the expected popularity of a video.

5.3 | Enhanced radio management system

5G is meant to support use cases where mission-critical communication, such as remote vehicle control, is done through the 5G network. These types of use cases require very high reliability from the network, as delay or loss of information could lead to fatal accidents. One of the main causes of service degradation in mobile networks are shadowing effects, which occur from the movement of the UEs and unfavorable conditions in their environment. These degradations need to be prevented or accounted for to achieve high reliability. Currently, much of the necessary knowledge (prediction of UE movements and radio conditions) resides in the management domain. Hence, the RAN domain cannot use this information to rapidly adjust the configuration of the radio link.

For improving radio coverage, 5G-NR supports various forms of beamforming. This capability allows the radio network to dynamically change the radio coverage to focus on hotspots, or low visibility areas in order to boost signal strength, overcoming interference or environmental shadowing effects. In order to combine management-domain analytics capabilities with (radio) network-domain beam control capabilities, we propose here an enhanced radio management system that considers a few targeted high-importance users, whose radio metrics are exposed to the management domain. In the management domain, an AI-based module performs medium to long term prediction of the radio conditions for each user, that are hence exposed back to the network function domain (the radio network, in this case) to perform beamforming reconfiguration and proactively mitigate shadowing effects.

In the context of the proposed architectural framework, the following capabilities are exposed:

- Producing signal strength data (capability type 1): This capability is provided and registered by the (radio) access network function domain (steps 1-4 in Figure 5). In this use case, it is discovered and consumed by an AI module in the management domain (steps 5-12).
- Performing of medium- to long-term prediction of radio conditions (capability type 4): This capability is produced and registered by two AI modules in the management domain (steps 1-4). While the first module implements mobility pattern prediction (MPP), the second one implements a network simulator (NS). In the example of the seaport use case, the radio network control entity responsible for beam control acts as the CCF and hence needs to discover and consume the output of the AI modules (steps 5-12).
- Re-configuring radio beams (capability type 3): This capability is produced and registered by network elements (ie, gNBs) of the (radio) access network function domain (steps 1-4). For the realization of the seaport coverage optimization use case, the radio network control function must discover and consume this capability (steps 5, 8-12). Since both CPF and CCF reside in the RAN domain, this is a simplified case of intra-domain capability exposure and steps 6 and 7 can be omitted.

6 | ARCHITECTURAL IMPLEMENTATION

In this Section, we provide the details about the implementation of one of the previously discussed use cases (the enhanced radio management system introduced in Section 5). Thus, here we discuss the different implementation steps of the Predictive Location-Aware Network Automation for Radio management (PLANAR), that effectively embodies the exposure functionalities discussed in Section 5.3.

With this discussion, we demonstrate the feasibility of our proposed architecture and the advantages brought by enabling new network exposure functions. In particular, many of the functions that are needed to implement PLANAR, are currently not supported by state-of-the-art solutions. Through discussing PLANAR, we provide a proof of concept to our proposed data exposure framework, highlighting where the new proposed methods take place . In the following, we also detail the implementation strategies of PLANAR and its integration into a real-world testbed. We remark that PLANAR has been implemented on a precommercial, although large-scale system, and thus some of the interfaces that we discuss in, for example, Section 4, could not be implemented in their entirety. Still, PLANAR serves as a feasibility proof

of concept for our system, and provides a qualitative feasibility analysis for the service based network exposure capability architecture proposed in this article. The interested reader may find more details about the testbed in Reference 42.

6.1 | The PLANAR testbed

The PLANAR concept is envisioned to be used in a smaller scale, privately-owned campus networks, such as factories, however many of the used interfaces can be extended to support any kind of Non Public Network (NPN) scenario. In NPN scenarios, the creation of a digital twin of the network (such as the one deployed for PLANAR) is likely possible, as the network is deployed in a limited area with available detailed information about the layout of the environment (such as a digital 3D floorplan). However, a specificity of the PLANAR framework is the one related to coverage, as areas with limited coverage are not easily fixed by additional cells, because of cost constraints, interference considerations, or dynamically changing environments, with shadowing objects moving around.

Consequently, the PLANAR concept has been studied in cooperation with the Hamburg Port Authority (HPA), implementing a fully operational 5G Network in the port of Hamburg that provided different industry-relevant use cases using the network slicing concept. These use cases are the following:

- **AR (augmented reality)**: Dock workers use AR goggles, which supply them with context-dependent information about their tasks on the fly. This application requires a relatively high bandwidth but does not have strict requirements with regards to reliability and latency. The AR network service is attached to the extreme/enhanced Mobile BroadBand (eMBB) slice.
- **Traffic light control**: This use case involves the remote control of mobile, deployable traffic lights in the port. The application does not consume a lot of bandwidth, but is sensitive to latency and requires a reliable connection. This application is attached to the Ultra-Reliable Low-Latency Communication (URLLC) slice.
- **Emission sensor readings**: In this use case, emission sensor (CO2) readings from port-maintenance barges use the mobile network. The readings have to be reliably transferred to the HPA control center to be able to intervene if a nearby ship was producing unacceptable amounts of emission. This application is attached to the Internet of Things (IoT) slice, and it is the use case we focused on when evaluating PLANAR.

The testbed consisted of two cells, both housed in the Heinrich Hertz tower, overlooking the port of Hamburg, each using four beams (as depicted in Figure 6). Because the cells were covering the relatively large area from a single direction, some sections of the port's waterways were occluded by large buildings, cranes, or the terrain. When barges ventured into these areas, emission sensor readings would often be delayed or completely lost.

FIGURE 6 The PLANAR testbed in Hamburg, Germany

6.2 | PLANAR implementation

Before discussing the need for enhanced network exposure, we first introduce the implementation of PLANAR from a system-level perspective, discussing the algorithmic choices behind it. The PLANAR system relies on an AI module that collects, merges, and analyzes data from different domains in order to predict the above-mentioned service impairments, such as radio quality indicators and position of the terminals. As a matter of fact, prior research involving the prediction of RSRP (reference signal received power) or other radio quality KPIs approach this problem as a time-series prediction. Works such as References 43 and 44 try to predict from highly granular measurements, frequently from a single domain, utilizing the same KPIs that were predicted without additional contextual information (such as location). Since the radio quality is subject to rapid changes both in time and in space, these time-series prediction-based approaches are only able to predict radio quality for a short timeframe forward, usually on the scale of a few hundred milliseconds to a few seconds. This short timeframe is not enough to undertake the larger-scale network reconfigurations that we envisioned for PLANAR, such as power control and beam re-configuration.

For longer predictions, the predictive system has to take into account data that allows the isolation of the noise-like variance from the actual cause of the radio quality changes. As mentioned earlier, the biggest cause of radio quality degradation is environmental shadowing,[45] which occurs when the UEs move behind large objects. Longer-term prediction of radio link quality can be achieved by splitting the task: first, the location of the users is predicted precisely far in the future and using this predicted location the expected radio quality is inferred through a location-dependent model of the radio environment.

6.2.1 | Data and exposure

Although 5G specifications (3GPP Rel.15 and later) include a variety of user localization techniques, many of the more precise ones require the explicit involvement of a system outside the network operator domains in order to achieve a good level of accuracy. GNSS-based (global navigation satellite systems), IMU-based (inertial measurement unit) or TBS-based (terrestrial beacon system) positioning systems comprise third party services allowing the terminal to provide additional measurements (from the service provider domain) to the network or the management domain. While this co-operation cannot be taken as granted from generic subscribers, it is the case in privately-owned campus networks and the associated subscribers such as this one.

For our implementation, different types of data have been collected from different domains of the testbed, which included UE-specific service-provider-level logs of GPS (global positioning system) coordinates, and ping, as well as radio quality measurements in the form of RSRP and RSRQ (reference signal received quality) values available in the RAN domain. Further, cell- and slice-specific KPIs were logged, such as throughput or PRB (physical resource block) utilization. The data collection took place over a time of six months. Both the sub-second granularity RSRP and RSRQ measurements, as well as the 1-second granularity PRB usage KPIs were aggregated (averaged) to the 5-second granularity of the GPS coordinates. This averaging also removed a lot of variance from these measurements, which, apart from the normalization of the value ranges to 0-mean and 1-variance, did not necessitate any additional preprocessing (such as a moving average calculation). All in all, this collection resulted in around three million records. For the training of the MPP module, this dataset was split into 65 206 individual 40-step-long (5-second granularity) sequences, where the first 32 steps represented the immediate past, and the last 8 steps the to-be-predicted future measurements of the barges. In total, around three million records were collected.

6.2.2 | PLANAR machine learning

Using the recorded locations, an MPP (mobility pattern prediction) module was created using a deep CNN (convolutional neural network)[46] to predict the movement of the barges. The input to the MPP consisted of a fixed-length sequence of historical locations of one of the barges, up to the most recent location. The output was a fixed-length prediction of future locations the barge will visit. The deep CNN was able to learn context-dependent routes around the port, such as recognizing when a barge was aligning to the shore for docking, and correctly predicting its movement even in this unusual

situation (Figure 7). The MPP module is our own implementation, using the PyTorch library for the CNN implementation and hardware acceleration in the Python scripting language. The CNN is a 3-layer deep network, the architecture of which can be seen on Figure8[‖]

To model the radio quality in a location-specific way, a digital twin of the city of Hamburg was created, complete with the precise modeling of buildings (including cranes and other large-scale equipment) and terrain. Using this digital twin and the radio quality measurements the barges collected, a model was fine-tuned to be able to precisely recreate the radio environment of the whole port. The digital twin is used as most of the physical systems that are deployed in the harbor and its surroundings cannot be constantly polled and measured by our system (such as the orientation of cranes, which heavily impact the radio signal quality). So, PLANAR builds on this model to have a reliable feedback of the channel quality along time, in addition to the live measurements.

Both the MPP and the model were implemented as separated AI modules in the management domain (Figure 9). After their training phases, the MPP module fed predicted location information to the model (intra-domain consumption of capability type 4), which gave the predicted radio quality as output. The model output is consumed by "actors" for NF control (inter-domain exposure and consumption of capability type 4), which changed the beam configuration and transmission power settings in the cells. These control functions do not only take as input predicted radio conditions from the model, but also feed the executed re-configurations back to it. All the communication between the different modules (which go beyond the interfaces defined by major SDOs), were implemented with *ad-hoc* Python modules.

Before enforcing the configuration in the network, we perform several training loops in the model, to make sure that the taken decisions are optimal in the whole network. For this purpose, we are not only taking into consideration the target barges but other users or the requirements of other slices in the network. Because the radio quality predictions

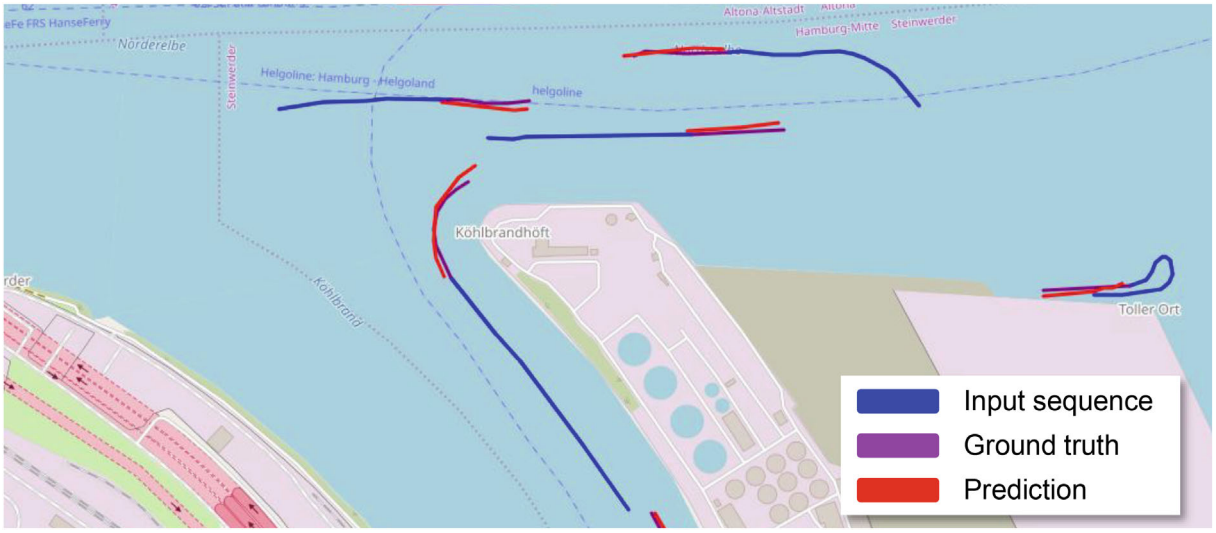

FIGURE 7 The barges' predicted mobility patterns

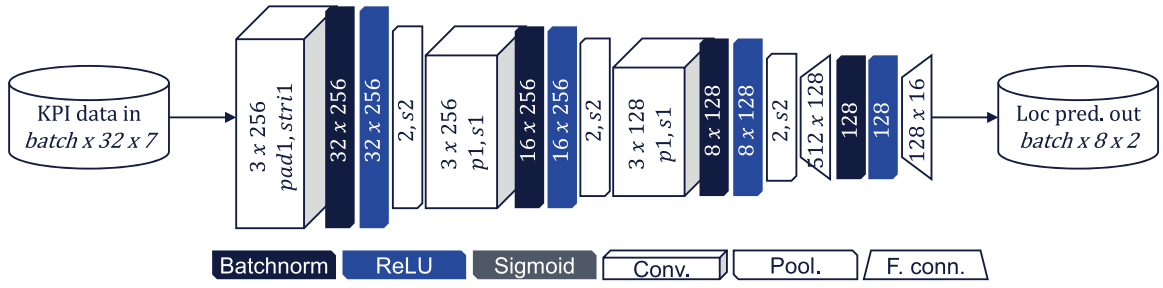

FIGURE 8 The deep learning topology used for the MPP in PLANAR

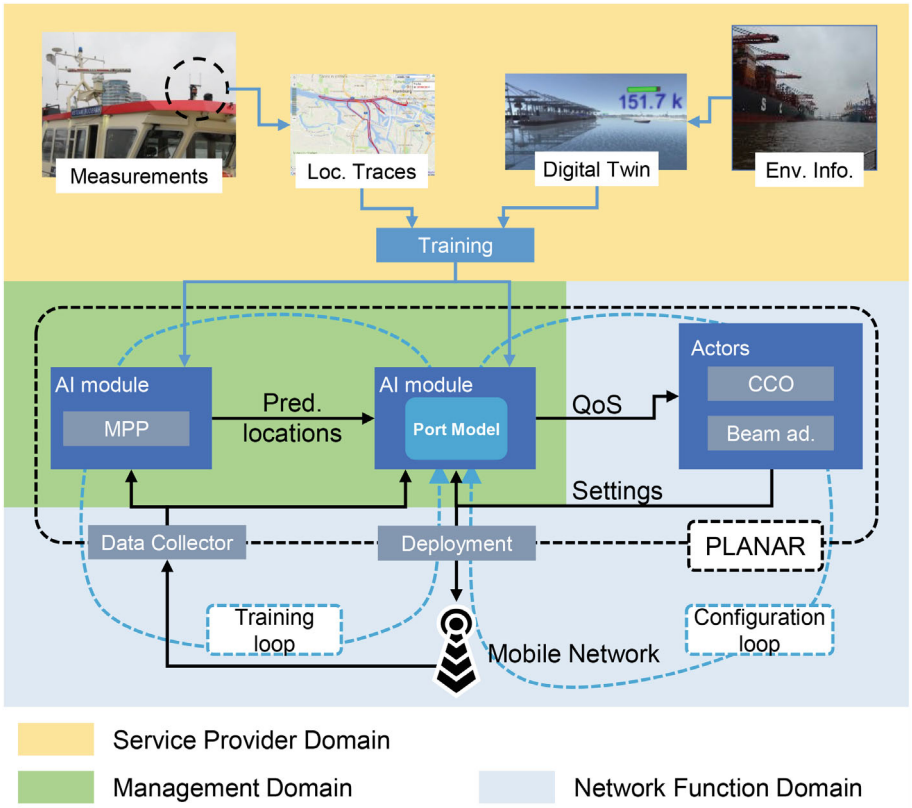

FIGURE 9 The PLANAR building blocks

were far ahead in the future, the actors had plenty of time to converge to a balanced configuration, at which point the configuration was deployed in the actual network.

Without realizing the capability exposure framework for the specific sea port setup, we could not have directly fed the PLANAR modules running in the management and NF domains, nor have interacted with the base stations deployed in the port. Without the secure, but open and uniform interfaces of the proposed framework, AI modules would have had to be placed in the service provider domain (ie, the port authority). The same would have been applied to the PLANAR Actors that control the beam adjustment. The chosen setup has allowed to include a broader set of data sources as well as a quicker and more effective reaction to events, as the individual control loop components could interact very seamlessly despite being placed in different domains: the PLANAR AI modules (MPP and the model) are directly deployed in the management domain and interact with the PLANAR actors as well as the data collectors implemented in the network functions domain).

While the implementation discussed in this article is limited in scope due to the constraints related to the real-world sea port testbed, the network exposure concept we discuss is widely applicable to any network deployment, including the public mobile networks providing, for example, network slicing capabilities.

6.3 | Evaluation results

As discussed before, we evaluated PLANAR taking into account a subset of barges location traces that were not used for training, which were fed to our system that

recreated the real network conditions for these traces and by predicted and avoided service degradation for the barges. Out of the 493 RLF (radio link failure) events that may have been taking place in our validation data, only 12 were not detected using the farthest (ie, most future-looking) predictions, achieving a 97.6% success rate. Later, as the barges got closer to the problematic areas, the incorrect predictions were also corrected, still before the actual RLF happened.

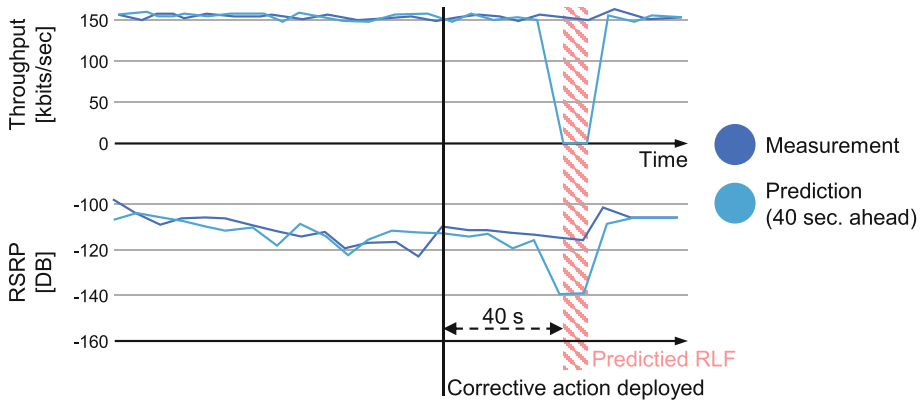

FIGURE 10 Example of a prevented radio link failure

TABLE 2 The capability producers and consumers in the PLANAR system

Producer (domain)	Consumer (domain)	Capability (type)
gNB (NFs domain)	PLANAR AI (Management domain)	Monitoring data (Type 1)
Port systems & sensors (Service provider domain)	PLANAR AI (Management domain)	Training data (Type 1 and 2)
PLANAR AI (Management domain)	PLANAR actor (NFs domain)	QoS target levels (Type 4)
gNB (NFs domain)	PLANAR actor (NFs domain)	Beam adjustment (Type 3)

In Figure 10, an example of a prevented RLF is shown. The dark blue lines depict the actual measurements, while the light blue lines depict the predicted values as forecast 40 seconds before. In this example, one of the barges was on a trajectory that led behind a steep riverbank, which could have caused a complete radio link failure. The ship's route was immediately accurately predicted by the MPP module. Using this prediction, the RLF was avoided by increasing the serving beam's power and down tilting the beam to better target the ship.

Conceptually, the use of data from multiple domains in two separate modules (MPP and model) for the prediction of future network states has strong benefits. The proposed unified capability registration, discovery, and exposure framework allows for a modular setup, where multiple data sources and NF actors can be present in the system (and dynamically added/removed/reconfigured), each optimizing network parameters through relatively simple logic. By trying out configurations in the simulated environment first, the NF actors can converge to a mutually optimal solution, avoiding transient states in the actual network. However, all of this is not possible without a flexible, controllable data exchange between the different modules residing in different domains, as well as data gathered from the different layers and slices of the network.

6.4 | Final considerations

The PLANAR system embodies the network capability exposure concepts discussed throughout this article, implementing the functions and procedures needed to enhance the state of the art technologies. PLANAR implements a subset of the full network exposure framework discussed in Section 3. Table 2 details the consumer and producer roles of the PLANAR components, depending on the capability needed for the use case. PLANAR eminently involves capability exposure at the management domain, as its goal is to optimize the long term behavior of the network, but also includes interactions

toward the Network Functions domain as well as the Service Provider domain, where the data provisioning capabilities are produced by the systems and sensors of the port authority.

More specifically, PLANAR uses capabilities of all kinds for its operation. The first kind of exposed data is the monitoring of the signal quality from the gNB. Exposing such information is fundamental for the training of the model and it is currently not possible with the state-of-the-art architecture, which only provides slow file-transfer based solutions for the exposure of such data. Instead, by directly exposing such data to the management functions, even public cloud based machine learning platforms could be used (although this is not the case for the PLANAR implementation).

Similar considerations apply also for the data coming from the sensors, which belong to the service provider (in this case the Port of Hamburg). With an over the top solution, with sensor data fully confined in the operator premises and SNR data only available to the operator the PLANAR solution could not have been possible, with a *lose-lose* situation: the service provider had an overall worse coverage in the area, with a possibly higher deployment cost for the network operator.

Finally, the decisions taken by the PLANAR system need to be enforced: this involves the effective exposure of the QoS analytics from the service provider to the network operator that, in turn, offers the needed interfaces to effectively inject such decisions in the network functions through proper APIs toward the RAN NFs.

7 | CONCLUSIONS

In this article, we presented a novel framework for the exposure of network capabilities from different domains (ie, network functions, orchestration, management, and service providers) to allow for the closed-loop automation of network services. Further, in this article, we presented a categorization of all the possible capabilities, grouping them by their characteristics, and proposed an architecture for their exposure. Our proposal, which extends the current standardization efforts, defines the relevant procedures for the capability consumption and it puts forth enhancements in the form of unified SBA, intra-domain and cross-domain Registration and Discovery Functions, as well as domain specific exposure functions and the definition of the essential procedures that allow for the interaction with other network elements.

We discussed how to apply our proposal to provide closed-loop automation to three innovative use cases, and implemented one of them. Our real-world testbed showed the advantages of the AI-based management of the Radio Access Network by leveraging the proposed framework.

ACKNOWLEDGMENTS

Part of this work was performed in the context of the H2020 5G-MoNArch project (grant agreement no. 761445). The work of Marco Gramaglia has been partially funded by the H2020 5G-TOURS project (grant agreement no. 856950), and by the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D projects 6G-CLARION-NFD, 6G-CLARION-OR, 6G-CLARION-SI, and 6G-CLARION-OE.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ENDNOTES

*https://www.nokia.com/networks/solutions/cloudband/.

†https://www.onap.org/.

‡https://www.magmacore.org/.

§https://www.acumos.org/.

¶https://www.lfedge.org/projects/akraino/.

#https://www.openapis.org/.

‖The label represent the different parameters of the different layers. For instance, the first convolutional layer contains 256 3-wide filters, with a padding and stride of 1. The number of neurons in each layer and the other deep learning parameters were calibrated through experimentation, finding a balance between the performance of the network and the time elapsed to train it, aiming for rapid experimentation. Beyond three layers, there was not any significant improvement in training accuracy (the mean squared error loss function is used), but rather a tendency to overfitting the data, hence leading to a worse prediction when deploying the system. The MPP module interfaces with the NS module through a simple JSON interface, through which the predicted locations of the barges are communicated. The MPP module uses GPUs (Graphical Processing Units) for hardware accelerated processing, which would likely be placed in an edge-cloud close to the cells using PLANAR. The

networking requirements for PLANAR are quite low-bandwidth, requiring only the stream of the barge locations on the input side, and outputting a stream of QoS predictions. Overall, more than 600 k sequences, encompassing 3 months of measurement were used for the overall training of PLANAR. The validation, instead, was performed for 10 days.

ORCID

Marco Gramaglia https://orcid.org/0000-0001-9494-1853

REFERENCES

1. Sciancalepore V, Mannweiler C, Yousaf FZ, et al. A future-proof architecture for management and orchestration of multi-domain NextGen networks. *IEEE Access*. 2019;7:79216-79232. doi:10.1109/ACCESS.2019.2923364
2. Bega D, Gramaglia M, Bernardos CCJ, Banchs A, Costa-Perez X. Toward the network of the future: from enabling technologies to 5G concepts. *Trans Emerg Telecommun Technol*. 2017;28(8):e3205. doi:10.1002/ett.3205
3. 3GPP TS23.501, system architecture for the 5G System (5GS) Rel. 16; 2020.
4. Gorski PL, Lo IL, Nguyen HV, Torkian DB. SOA-readiness of REST. In: Villari M, Zimmermann W, Lau K-K, eds. *Service-Oriented and Cloud Computing, Lecture Notes in Computer Science*. New York: Springer; 2014:81-92.
5. Marsch P, Bulakci Ö, Queseth O, Boldi M. E2E architecture; 2018:79-114.
6. Barakabitze AA, Ahmad A, Mijumbi R, Hines A. 5G network slicing using SDN and NFV: a survey of taxonomy, architectures and future challenges. *Comput Netw*. 2020;167:106984. doi:10.1016/j.comnet.2019.106984
7. Pateromichelakis E, Moggio F, Mannweiler C, et al. End-to-end data analytics framework for 5G architecture. *IEEE Access*. 2019;7:40295-40312.
8. Gutierrez-Estevez DM, Gramaglia M, De Domenico A, et al. Artificial intelligence for elastic management and orchestration of 5G networks. *IEEE Wirel Commun*. 2019;26(5):134-141. doi:10.1109/MWC.2019.1800498
9. Davoli G, Cerroni W, Tomovic S, Buratti C, Contoli C, Callegati F. Intent-based service management for heterogeneous software-defined infrastructure domains. *Int J Netw Manag*. 2019;29(1):e2051. doi:10.1002/nem.2051
10. Katsalis K, Nikaein N, Edmonds A. Multi-domain orchestration for NFV: challenges and research directions; 2016:189-195.
11. 3GPP TS23.002, network architecture Rel. 162020.
12. Jorguseski L, Pais A, Gunnarsson F, Centonza A, Willcock C. Self-organizing networks in 3GPP: standardization and future trends. *IEEE Commun Mag*. 2014;52(12):28-34. doi:10.1109/MCOM.2014.6979983
13. 3GPP TR28.801, telecommunication management study on management and orchestration of network slicing for next generation network Rel. 16; 2020.
14. 3GPP TS23.288, architecture enhancements for 5G System (5GS) to support network data analytics services Rel. 16; 2020.
15. Kukliński S, Tomaszewski L, Kołakowski R. On O-RAN, MEC, SON and network slicing integration; 2020:1-6.
16. 3GPP TS38.300, NR; overall description; stage-2 Rel. 16; 2020.
17. 3GPP TS38.401, NG-RAN; architecture description Rel. 16; 2020.
18. 3GPP TS29.531, 5G system; network slice selection services; stage 3 Rel. 16; 2020.
19. 3GPP TS23.502, Procedures for the 5G System Rel. 162020.
20. 5G-MoNArch D2.3, final overall architecture; 2019. https://5g-monarch.eu/wp-content/uploads/2019/05/5G-MoNArch_761445_D2.3_Final_overall_architecture_v1.0.pdf
21. 3GPP TS28.530, management and orchestration; concepts, use cases and requirements Rel. 16; 2020.
22. 3GPP TR28.809, study on enhancement of management data analytics Rel. 16; 2020.
23. ETSI, network functions virtualisation (NFV). Management and orchestration; Vol. 1, 2014:V1.
24. ETSI, OSM. Open Source Mano. OSM home page; 2020.
25. González S, Oliva A, Costa-Pérez X, et al. 5G-Crosshaul: an SDN/NFV control and data plane architecture for the 5G integrated Fronthaul/Backhaul. *Trans Emerg Telecommun Technol*. 2016;27(9):1196-1205. doi:10.1002/ett.3066
26. 5G-ACIA, 5G alliance for connected industries and automation, a working party of ZVEI (German electrical and electronic manufacturers' association). https://www.5g-acia.org/
27. Droste H, Rost P, Doll M, et al. An adaptive 5G multiservice and multitenant radio access network architecture. *Trans Emerg Telecommun Technol*. 2016;27(9):1262-1270. doi:10.1002/ett.3087
28. Trichias K, Landi G, Seder E, et al. VITAL-5G: innovative network applications (NetApps) support over 5G connectivity for the transport amp; logistics vertical; 2021:437-442.
29. Marsch P, Da Silva I, Bulakci Ö, Tesanovic M, El Ayoubi SE, Säily M. Emerging network architecture and functional design considerations for 5G radio access. *Trans Emerg Telecommun Technol*. 2016;27(9):1168-1177. doi:10.1002/ett.3073
30. Chih-Lin I, Kuklinskí S, Chen T. A perspective of O-RAN integration with MEC, SON, and network slicing in the 5G era. *IEEE Netw*. 2020;34(6):3-4. doi:10.1109/MNET.2020.9277891
31. Lin L, Zhu B, Wang Q, Xu L, Mu J. A novel 5G core network capability exposure method for telecom operator; 2020:1450-1454.
32. Ortiz J, Sanchez-Iborra R, Bernabe JB, Skarmeta A, INSPIRE-5Gplus: intelligent security and pervasive trust for 5G and beyond networks. ARES'20; 2020; Association for Computing Machinery, New York, NY.
33. Szabó G, Seres G, Mikecz ML, et al. Assessment of the efficiency of 5G network exposure for the industrial Internet of Things; 2021:52-58.
34. GSMA, generic network slice template version 1; 2019.

35. 3GPP TR23.700-40, Study on enhancement of network slicing; Phase 2 Rel. 17; 2020.
36. 3GPP TS28.541, management and orchestration; 5G network resource model (NRM); Stage 2 and stage 3 Rel. 16; 2020.
37. 3GPP TS23.755, Study on application layer support for Unmanned Aerial Systems (UAS) Rel. 172020.
38. 3GPP TS23.287, study on application support layer for factories of the future (FotF) in the 5G network Rel. 17; 2020.
39. OpenAirInterface: democratizing innovation in the 5G era. *Comput Netw*. 2020;176:107284. doi:10.1016/j.comnet.2020.107284
40. Benzaid C, Taleb T. AI-driven zero touch network and service management in 5G and beyond: challenges and research directions. *IEEE Netw*. 2020;34(2):186-194. doi:10.1109/MNET.001.1900252
41. Antevski K, Bernardos CJ, Cominardi L, Oliva A, Mourad A. On the integration of NFV and MEC technologies: architecture analysis and benefits for edge robotics. *Comput Netw*. 2020;175:107274. doi:10.1016/j.comnet.2020.107274
42. Rost P, Breitbach M, Roreger H, et al. Customized industrial networks: network slicing trial at Hamburg seaport. *IEEE Wirel Commun*. 2018;25(5):48-55. doi:10.1109/MWC.2018.1800045
43. Hongjia L, Song C, Zejue W. Prediction handover trigger scheme for reducing handover latency in two-tier Femtocell networks; 2012:5130-5135.
44. Malanchini I, Suryaprakash V. Minimizing the impact of prediction errors during anticipatory resource allocation; 2018:1-6.
45. Agrawal P, Patwari N. Correlated link shadow fading in multi-hop wireless networks. *IEEE Trans Wirel Commun*. 2009;8(8):4024-4036.
46. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks; 2012:1097-1105; Curran Associates, Inc.

How to cite this article: Gramaglia M, Kajo M, Mannweiler C, Bulakci Ö, Wei Q. A unified service-based capability exposure framework for closed-loop network automation. *Trans Emerging Tel Tech*. 2022;e4598. doi: 10.1002/ett.4598