5G Automation Architecture

White paper

Authors:
Milan Stolic
Julie Ann Connary
Wei Yan
Arghya Mukherjee
Mohammad Salaheldin
Rob Piasecki

Contents

Abstract

Mobile service providers are facing increased customer demand with stagnant average revenue per user (ARPU), forcing the need for new revenue-generating services. Network transformation is needed in all network domains to meet this demand and enable new revenue streams for service providers. This transformation remains daunting for many operators. [1]

5G by its nature spans many domains beyond mobility and requires a transformation unseen so far. This paper will provide a simplified view of 5G automation architecture by focusing on per-domain and cross-domain automation and orchestration, while taking into consideration 5G components such as Cloud-RAN, Control User-Plane Separation (CUPS), Multiaccess Edge Computing (MEC), network slicing, and xHaul that have to work in tight synergy.

Scope

This document provides a reference for 5G network automation and orchestration in general terms, applicable to most use-case categories. It is an architecture guidance that glues the different components of 5G automation together. This paper can be shared internally within Cisco and externally.

This document is targeting managers, program managers, and planning, implementation, and operations engineers on both SP and vendors' sides involved in 5G automation planning and implementation. It provides an at-a-glance view of options and considerations for orchestrating different aspects and domains of a 5G network.

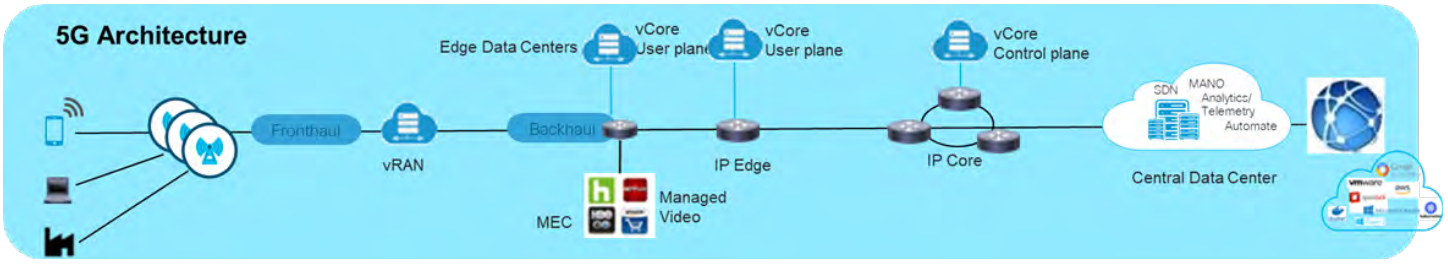
Introduction

5G is a fundamental transformation of mobile technology that will lay the path for many future technologies. 5G positions mobile technology as the core platform for advancing the innovations in future defining technologies [1].

5G enables service providers to be an integral part of various businesses, driving growth by providing customized services beyond connectivity. This is very important for SPs constantly struggling with identifying new revenue opportunities.

At a high level, the 5G network architecture is depicted in the figure below. Even at a high level there are significant additional complexities compared to the previous generations: it is access agnostic and has disaggregated Radio Access Network (RAN); fronthaul and/or midhaul are introduced; Mobile Core has new elements; Control and User planes are separated; and edge computing is introduced, to name a few. Understanding evolving requirements, stitching together the pieces of the new network, and transitioning services and traffic from a very different existing network can be a complicated task.

Figure 1. 5G architecture



As with many other major changes in the industry, 5G is expected to arrive in phases with a significant impact on existing infrastructure. This document will not focus on deployment phases including 5G NSA (non-standalone); its focus will be 5G automation and orchestration architecture in terms of functionality, and with respect to different network domains and cross-domain, end to end. It will also provide a view and mapping to the 3GPP and ETSI standards, as they apply to the Automation Architecture.

3.1 Standards view

Several standard bodies and working groups are actively identifying key requirements and interfaces required for 5G network automation. These different standard bodies and working groups are complementing each other in order to identify an end-to-end orchestration strategy.

The key standard bodies and working groups are:

- 3GPP
- TMForum
- ETSI NFV
- ETSI ZSM
- ONAP

In the next chapter, we define the key roles and outcomes from each of the standard bodies.

3.1.1 3GPP

3GPP is concerned with identifying a management and orchestration architecture for 3GPP domains, namely 5G core and RAN, in order to achieve the main use case for automation, which is 5G slicing. 3GPP has defined the architecture in 3GPP standards TS 28.530 [4], TS 28.531 [5], TS 28.532 [6], and TS28.533 [7].

3GPP identified the following main network functions to aid the 5G automation:

- CSMF: Communication Service Management Function
- NSMF: Network Slice Management Function
- NSSMF: Network Sub-Slice Management Function
- NFMF: Network Function Management Function
- MDAF: Management and Data Analytics Function
- NWDAF: Network Data Analytics Function

3.1.2 TMForum

The TMForum provides an end-to-end reference framework for automation and orchestration across multiple domains that includes standard interfaces for integration with Operations and Billing Support Systems (OSS/BSS).

In general, the TMForum framework provides technology standard body information models when they are needed. For 5G network slicing, it provides alternative specifications to the 3GPP for interfaces between technology domains.

The TMForum OpenAPI work identifies the following key standard interfaces that can be used for OSS/BSS integration to a network domain:

- TMF640/641 Service Activation and Configuration API [9][10]
- TMF638 Service Inventory API [11]
- TMF633 Service Catalogue API [12]
- TMF639 Resource Inventory Management API [23]

3.1.3 ETSI NFV

ETSI NFV (Network Function Virtualization) identifies the NF (Network Function) lifecycle management procedures in a virtualized environment.

Our focus in this whitepaper is a part of the ETSI standard specifying the procedure needed to integrate with the 5G management and orchestration framework (especially at the NSSMF–NFVO connection and NFMF–VNFM level, as defined in 3GPP standards) for slice instantiation and lifecycle management of VNFs and CNFs.

ETSI specifies interfaces in ETSI NFV standards, namely the SOL005 [8] interface, where the NSSMF can be used to create, scale, or delete a virtualized network function.

3.1.4 ETSI ZSM (Zero-Touch Network and Service Management)

The reference architecture—defined in ETSI GS ZSM 002 V1.1.1 [13]—employs a set of architectural principles and a service-centric architectural model to define, at a high level, a set of management services for zero-touch network and service management. It also defines a means of management service integration, communication, interoperation, and organization at a functional level. Procedures and detailed information models are beyond the scope of the present document. The reference architecture also defines normative provisions for externally visible management services, defined as part of the reference architecture, as well as recommendations for their organization. It is assumed that the architectural patterns introduced in the present document can be used not only for the ZSM framework, but also for architecture and design of individual management services.

ETSI ZSM has 12 design principles:

- Modularity
- Extensibility
- Scalability
- Model-driven open interfaces
- Closed-loop management automation
- Stateless management functions
- Resilience
- Separation of concerns in management
- Service composability
- Intent-based interfaces
- Function abstraction
- Simplicity

To achieve the purpose of the design principles, the architecture was developed with the following building blocks:

- Management services
- Management functions
- Management domains
- E2E service management domain
- Integration fabric
- Data services

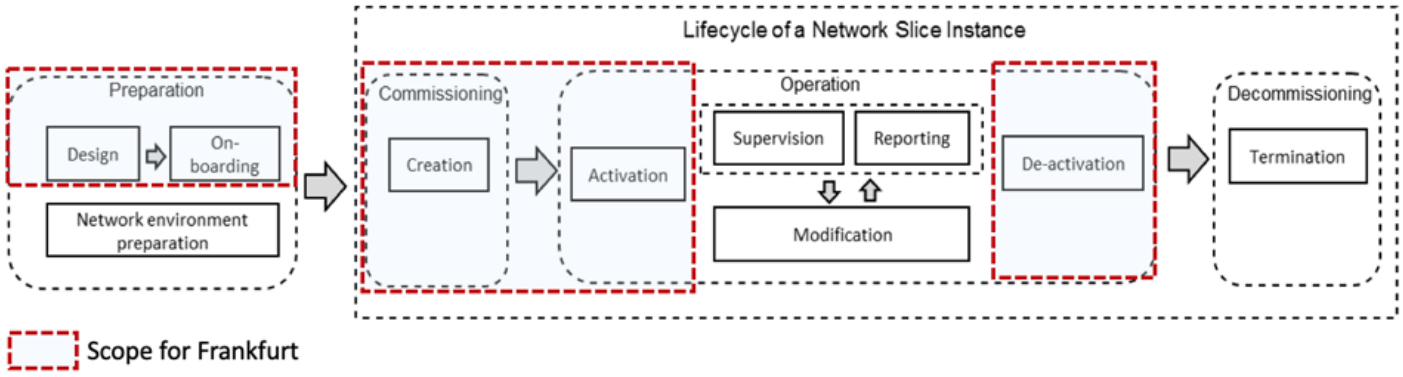
3.1.5 The Linux Foundation–ONAP Project

ONAP[14] defined network slicing as a key use case in the required service provider automation. ONAP recognizes this effort as a multi-release effort to achieve the automation use cases.

At the time of writing this document, ONAP's next release is Frankfurt release [15][16], which specifies implementing only CSMF and NSMF functions as part of ONAP, and integrating with a third-party NSSMF using standard APIs.

From the Network Slice Instance (NSI) lifecycle management point of view, in this release, ONAP will implement functions in the red box below, which includes NSI design and preprovisioning, NSI instantiation and configuration, and NSI activation and deactivation.

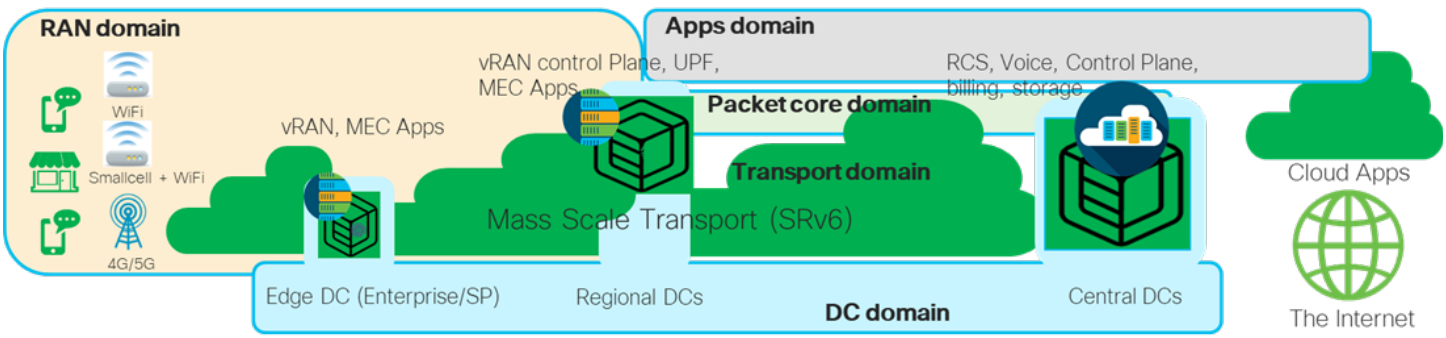
Figure 2. Scope for ONAP Frankfurt release



3.2 5G domains

While there is no strict definition of a "domain," a service provider's networks can be loosely divided into the following domains: RAN, Transport, Packet Core, Data Center (DC), and Application. This division is not strictly physical; it looks at different functional blocks used to build a network, and these blocks may physically overlap. However, logical division is important from the automation and orchestration perspective. Details and configuration of each domain's elements will vary depending on the specific use case [2]. A high-level diagram of network domains, as described, is shown below.

Figure 3. 5G network domains



3.3 Network slicing

Among the 3GPP specifications for the 5G core system are three key types of network services driving the need for network slicing, also viewed as use-case categories:

- Enhanced Mobile Broadband (eMBB): Requirements focused on greater bandwidth and moderate improvements to latency for 4G LTE and 5G NR deployments.
- Ultra-Reliable Low-Latency Communications (URLLC): Provides increased bandwidth for 5G core deployment with a focus on end-to-end latency reduction.
- Massive Machine-Type Communications (mMTC): Developed to provide connectivity to a larger set of devices (for example, Internet of Things [IoT] sensors) that typically transmit small blocks of data via low-bandwidth paths.

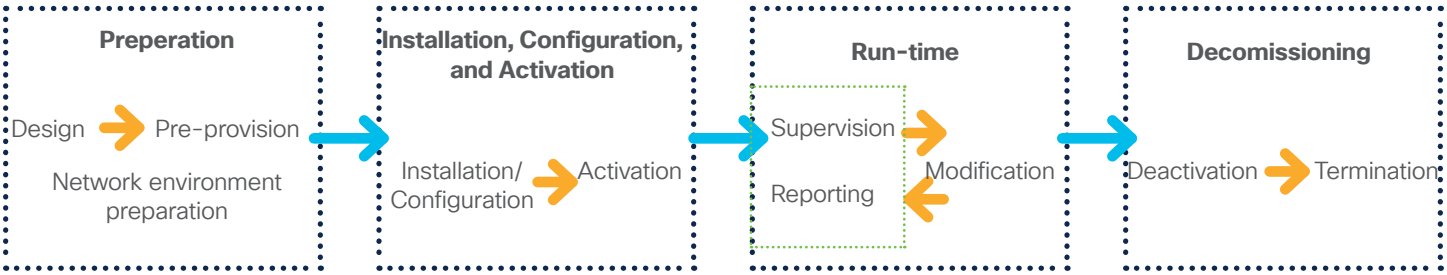
The network slicing concept consists of three layers [3]:

1. Service Instance Layer,
2. Network Slice Instance Layer, and
3. Resource Layer

Each service is represented by a Service Instance. Typically, services can be provided by the network operator or by third parties. A Service Instance is an instance of an end-user service or a business service that is realized within or by a Network Slice.

A Network Slice Instance (NSI) is a managed entity in the operator's network with a lifecycle independent of the lifecycle of the service instance(s). In particular, service instances are not necessarily active through the whole duration of the run-time phase of the supporting NSI [3].

Figure 4. NSI lifecycle



The following phases describe the network slice lifecycle [3]:

- **Preparation phase:** Creation and verification of network slice template(s), onboarding, preparing the necessary network environment used to support the lifecycle of NSIs, and any other preparations needed in the network.
- **Instantiation, Configuration, and Activation phase:** All resources shared/dedicated to the NSI have been created and are configured (such as, to a state where the NSI is ready for operation).
- **Run-Time phase:** NSI is capable of traffic handling to support communication services of a certain type(s). This phase includes monitoring, as well as activities related to modification. Modification could map to several workflows related to run-time tasks (for example, upgrade, reconfiguration, NSI scaling, changes of NSI capacity, changes of NSI topology, and association and disassociation of network functions with NSI).
- **Decommissioning phase:** Deactivation, the reclamation of dedicated resources, and configuration of shared/dependent resources.

3.4 Necessity of automation

Although at this point 5G is predominantly in initial testing and deployment phases, most MNOs have not rushed to implement automation; however, they mostly understand the associated challenges [18]. The number of network elements needed to run a 5G network in all domains means that any cost-effective 5G deployment will require automation to deploy efficiently, and keep things running. The reason for hesitation is a perceived lack of maturity and slow progress in automation technology.

In an automated deployment scenario, all or most of the heavy preplanning manual work can be eliminated. Artificial Intelligence (AI) systems, based on Machine Learning (ML), can model how network functions would perform under normal and high-stress conditions. Using run-time performance data, the system can ensure automatic deployment of new elements as needed in a Continual Integration/Continual Deployment (CI/CD) mode. For ongoing optimization and Service Assurance, systems can collect and analyze equipment feeds of all types and examine their performance, determining if it matches the parameters that SPs require and expect, as well as the customer experience of the end users.

3.5 Challenges of moving to the cloud

5G networks, including 5G core and edge, can be hosted in a public or private cloud, in a centralized or decentralized manner. A 5G network can be hosted entirely on a customer premise, or in an edge DC providing improvements in latency, availability, security, etc. [19].

However, hosting the application in the cloud comes with its own challenges:

- New operations management and security solutions are required
- Finding use cases and business models behind the cloud edge
- Clouds must support the required high throughput
- Operations, processes, security, and availability must meet the expectations of SPs and their customers

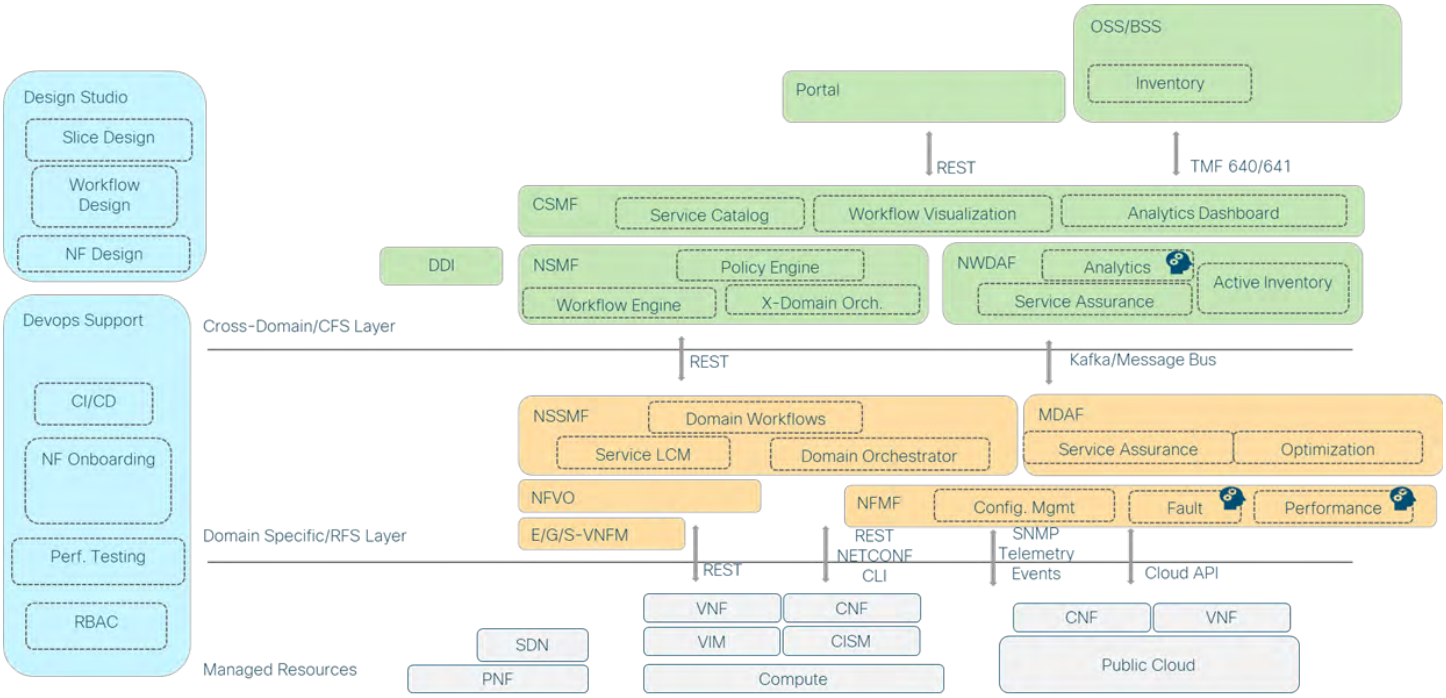
Cloud providers offer their own solutions to ease the design of moving services to the cloud [24][25].

Disaggregation of functionality is typical in future networks. Some parts of the infrastructure (for example, User Plane Function [UPF] or Virtual Centralized Unit [vCU]) may need to be co-located with the Application Function (AF), and applications served from the Cloud. With this new disaggregation comes additional parties to be involved, with corresponding clarifications about ownership, responsibilities, and management, that need to be made in a timely manner.

Solution overview

The broad scope of 5G automation requires a product-agnostic approach to a 5G system architecture that distills the key recommendations and principles of the various extant standards into a deployable operational framework. The solution shown below breaks down the various functions in the architecture in a way that is prescriptive enough to clearly define the necessary building blocks, while retaining the flexibility to work with the operational models and needs of all types of mobile service providers.

Figure 5. 5G automation architecture



The figure above shows a high-level overall 5G automation architecture that can be divided into the following areas:

- A cross-domain layer that acts across multiple domains while also presenting the majority of customer-facing services
- Domain-specific layers that translate the intent of the cross-domain layer into the domain intelligence that manages the appropriate resource-facing services
- Managed resources, which include virtualized, cloud-native, and physical network functions specific to a domain
- DevOps support functions essential for automation and operation

Functional elements of each layer are shown in the diagram above.

4.1 Customer-facing service and resource-facing service

The fundamental goals of a successful 5G automation architecture are to manage complexity and increase operational efficiency. A layered architecture that separates the specifics of the resource-facing services at the domain level from the customer-facing services at the cross-domain level is critical in fulfilling these goals. Key considerations for this separation are:

- Consistent abstractions
- Standardized interfaces
- Idempotent/reversible operations
- Clear distribution of control between domains

The intent is to allow the resource-facing functions greater control of fine-grained operations at the domain level, while presenting abstracted operations for the customer-facing functions to compose slice-level operations from multiple domains.

4.1.1 Functions in CFS and RFS

The following functions are common in the RFS layer in all the domains:

- NFVO and G-VNFM: Manage the lifecycle of virtualized or containerized workloads.
- NFMF: Manages configuration, fault, and performance of one or more individual network functions.
- NSSMF: Manages workflows and orchestrates and performs service lifecycle management at a domain level. It may also present some visualization capability such as topology.
- MDAF: Provides analytics and assurance capabilities at the domain level. Active inventory—the domain view of the state of all managed devices in order to provide correlation and running data—shall also be maintained while being synchronized with the overall static and physical inventory maintained by the OSS/BSS. It may also provide some domain-level optimization functions.

The following functions are present at the cross-domain/CFS layer:

- NSMF: Performs cross-domain slice management functions, utilizing a workflow engine and a cross-domain orchestrator. Performs policy enforcement and resource enforcement for all slice operations.
- NWDAF (Network Data Analytics Function): Provides analytics and assurance capabilities at the cross-domain (slice and service) level, while maintaining cross-domain active inventory. It will enable service-level, closed-loop operations via E-W integration with the NSMF.
- CSMF: Acts as a presentation layer—presents the service catalog, workflow visualization, and analytics/service assurance dashboard.

4.2 DevOps support functions

DevOps methodologies and their aligned support functions will be essential to deploy and operate the automation framework described above. These functions are the glue that hold the framework together to provide the logistics and command-and-control functions that enable all disparate components and humans in the system to interact in a consistent, secure manner.

- CI/CD: Supports the ability to rapidly test and deploy software and other network artifacts through the system.
- NF Onboarding: Supports the ability to rapidly add an NF or application to the service catalog.
- AAA: Access, authorization, and accounting framework that the entire automation architecture will use—includes certificate servers, RBAC, LDAP integration, audit mechanisms, and SOC integration.

Cross-domain orchestration

To achieve the goals of 5G automation, a cross-domain orchestration is needed to connect the parts together between different domains composing the network.

The cross-domain orchestrator, residing in the CFS layer of the solution, can receive service instantiation requests from the OSS/BSS, or a self-service portal (such as CSMF), and it works on fulfilling the service request in addition to other functions.

The cross-domain orchestrator is expected to interface with different domain orchestrators (DOs) residing in the domain-specific RFS layer in order to achieve the above task.

5.1 Cross-domain orchestration mapping to the standards

The following is the map where cross-domain orchestrator fits in the architecture provided by the different standard bodies:

- **TMF:** Cross-domain orchestrator fits in the E2E service orchestrator layer
- **3GPP:** Cross-domain orchestrator covers the functional requirements of CSMF, NSMF, and NWDAF
- **ETSI ZSM:** E2E service management domain
- **ONAP deployment:** ONAP intends to cover the cross-domain orchestration layer in its next architecture release.

5.2 Cross-domain orchestrator requirements

The cross-domain orchestrator provides multiple functions that are crucial for E2E service delivery and E2E closed-loop automation (zero-touch network).

The cross-domain orchestrator provides the following functions:

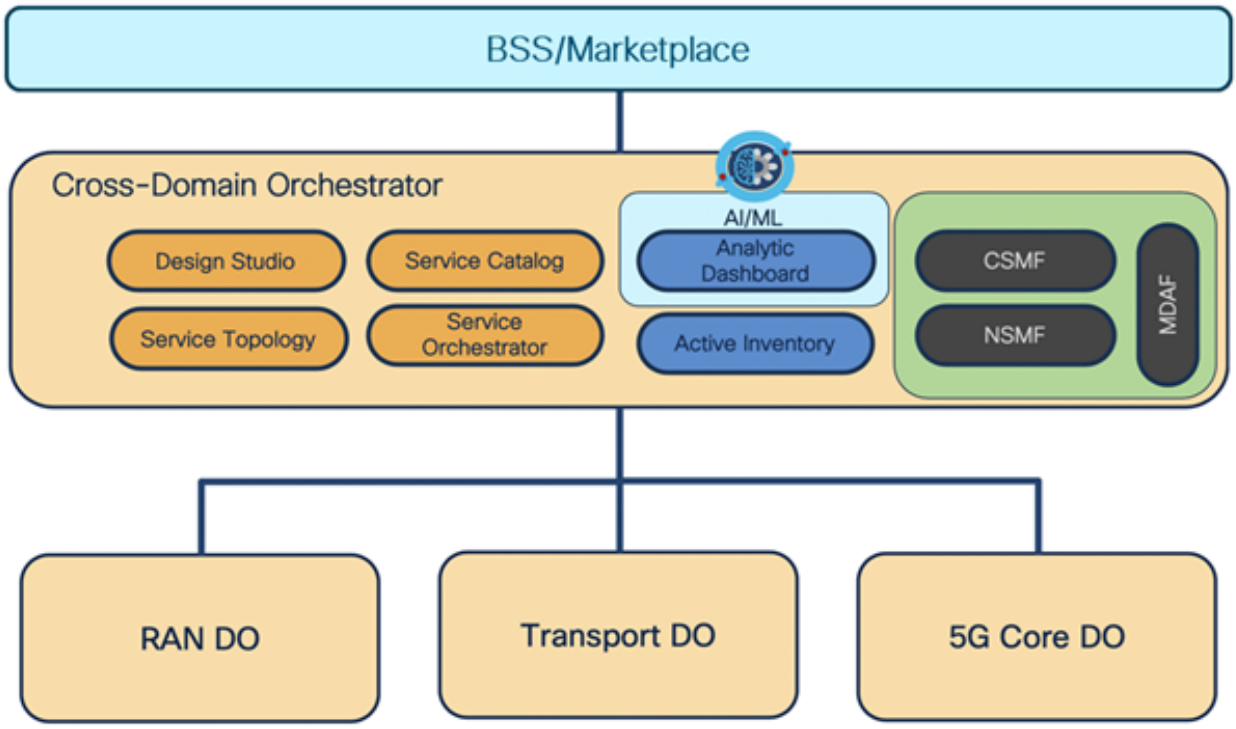
- Integrates with the OSS/BSS layer, receives order management requests (TMF 640/641) and exports available services on the service catalog to the BSS digital marketplace (TMF 633)—in case TMF interfaces are not ready on OSS/BSS and the REST API integration is provided.
- The service fulfilment module translates the order request into fulfillment logic, including pre-checks, service breakdowns, policies, and post-checks.
- Integrates with the RFS layer and possibly different domain orchestrators to fulfill the request on the resource level
- Service inventory to list all the instantiated services
- Service topology view
- Service design studio is provided to build the service creation logic using GUI.
- Service analytics module receiving telemetry and performance information from various RFS layers to consolidate KPI management, perform advanced analytics, and apply logic for closed-loop automation
- Closed-loop automation for service remediation

Note: The self-service portal in this solution is not a replacement for the self-service portal on the BSS layer (digital marketplace), which provides integration with the billing system, CRM, or any other required business support system.

5.3 Northbound and southbound integration

As illustrated in the diagram below, several integrations are needed to fulfill the cross-domain orchestration functionality.

Figure 6. Cross-domain orchestrator integration points



5.3.1 Northbound integration

An interface is required to communicate with:

- Service activation and configuration
- Service catalog
- Service inventory

Either general REST APIs are used with the BSS/marketplace or TMF standard APIs can be used, namely:

- TMF640/641 Service Activation and Configuration API [9][10]
- TMF638 Service Inventory API [11]
- TMF633 Service Catalogue API [12]

5.3.2 Southbound integration

Integration with different domain controllers is usually achieved using NETCONF or REST APIs.

If a direct integration with NFVO is needed, the SOL005 interface shall be used.

It is worth mentioning that 3GPP is in the process of defining specific standard APIs for integration with different RAN NSSMF and 5G core NSSMF in releases 16 and 17.

Domain-level orchestration

6.1 RAN domain orchestration

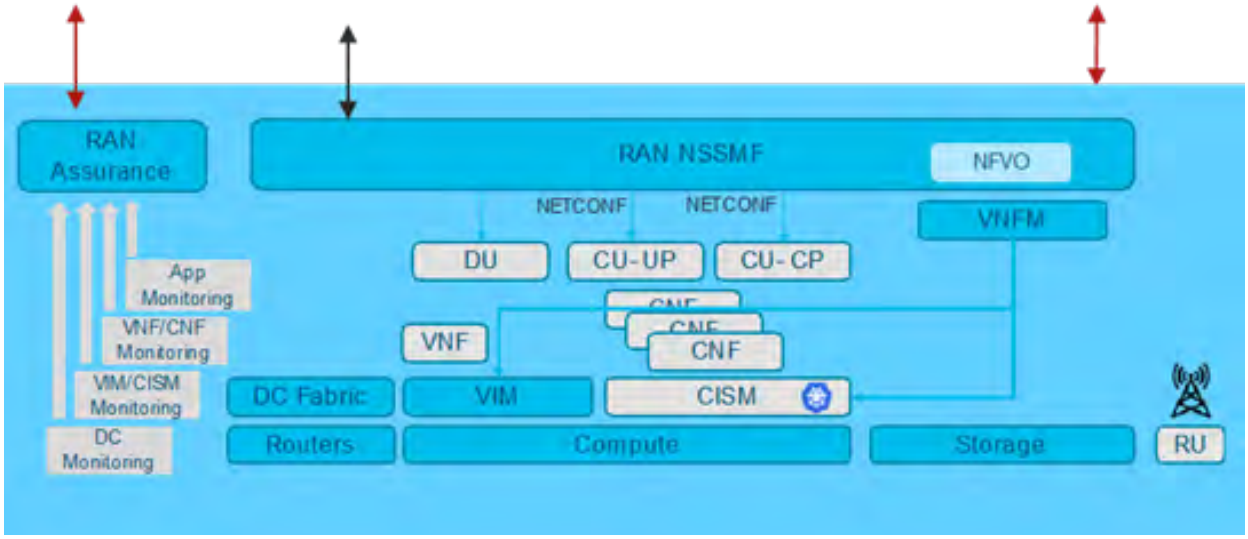
ITU R-M.2083 [27] describes the following use cases for RAN in 5G:

- Enhanced Mobile Broadband (eMBB): Addressing human-centric use cases for access to multimedia content requiring seamless coverage and medium to high mobility with much improved user data rates compared to existing (4G) data rates
- Ultra-Reliable Low-Latency Communications (URLLC): Addressing applications with stringent requirements of throughput, latency, and availability (such as industrial manufacturing and remote medical surgery)
- Massive Machine-Type Communications (mMTC): Addressing applications characterized by a very large number of connected devices transmitting a low volume of non-delay-sensitive data

These use cases require a high degree of automation at a scale that can be challenging. In addition, disaggregation into edge data centers and virtualization of the RAN are also critical themes that 5G brings to the fore. The NSSMF in this case will be required to support both virtualized and containerized workloads via NFVO and VNFM, while also managing some physical network functions (PNFs) such as cell site routers, microwave radio links, cell site switches, and radio interface units (RIUs).

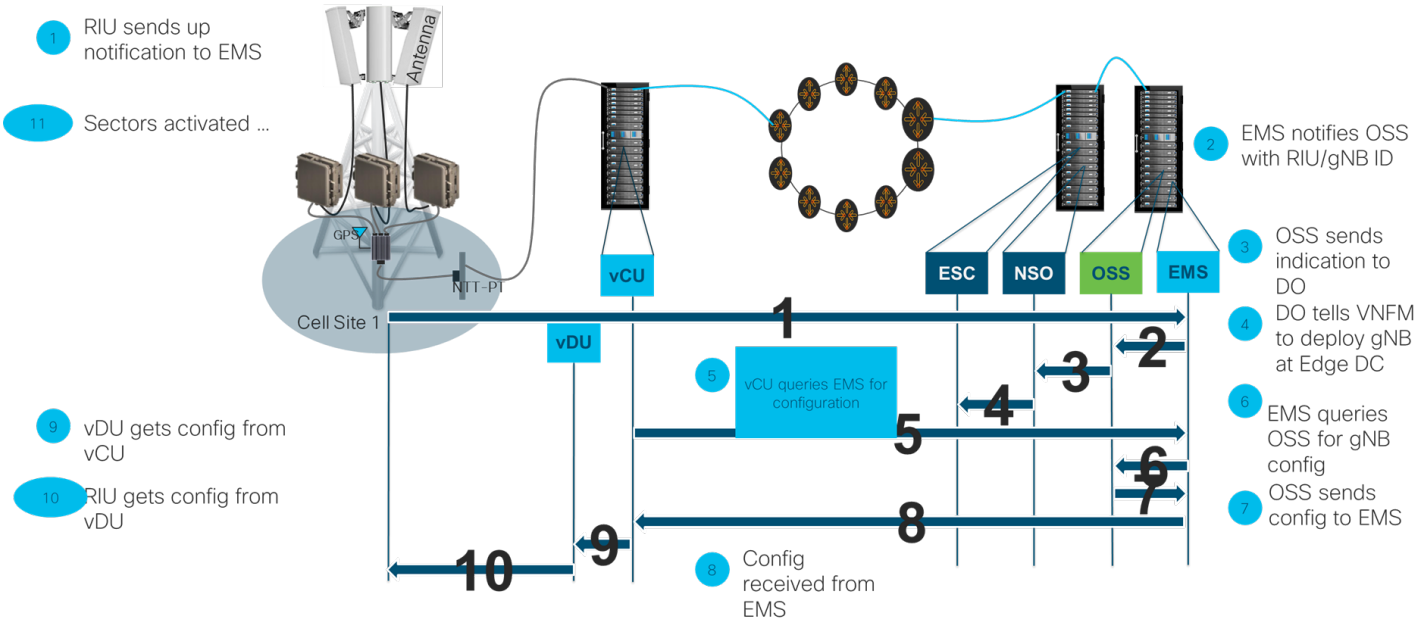
This architecture will be aligned as well with O-RAN specifications [20] for non-real-time control and other service provisioning activities.

Figure 7. RAN domain automation architecture



One example of a vRAN automation call flow is shown in the diagram below. It follows the key ETSI ZSM principles listed in section 3.1.4 of this document.

Figure 8. RAN domain automation call flow



In this real-life example, applicable to both 4G and 5G deployments in a virtualized environment, a new cell site is being activated, with the radio interface unit (RIU), vDU, and vCU configured automatically.

In addition to the site initialization, specific slice scheduling configuration needs to be managed by the RAN NSSMF in order to achieve the strict requirements of each and every slice.

In order to create a new slice, NSSMF performs the following:

- Upon receiving a Network Slice Subnet Instance (NSSI) instantiation request, pre-checks are done to make sure the current NSSI will not be able to support the new request.
- The available RAN resources are checked for availability of provisioning the new scheduling rules.
- The required scheduling information is configured for each of the RAN slices from existing templates (NSMT)
- The configuration requirements are sent to RAN NFMF for configuration management and further for service assurance.

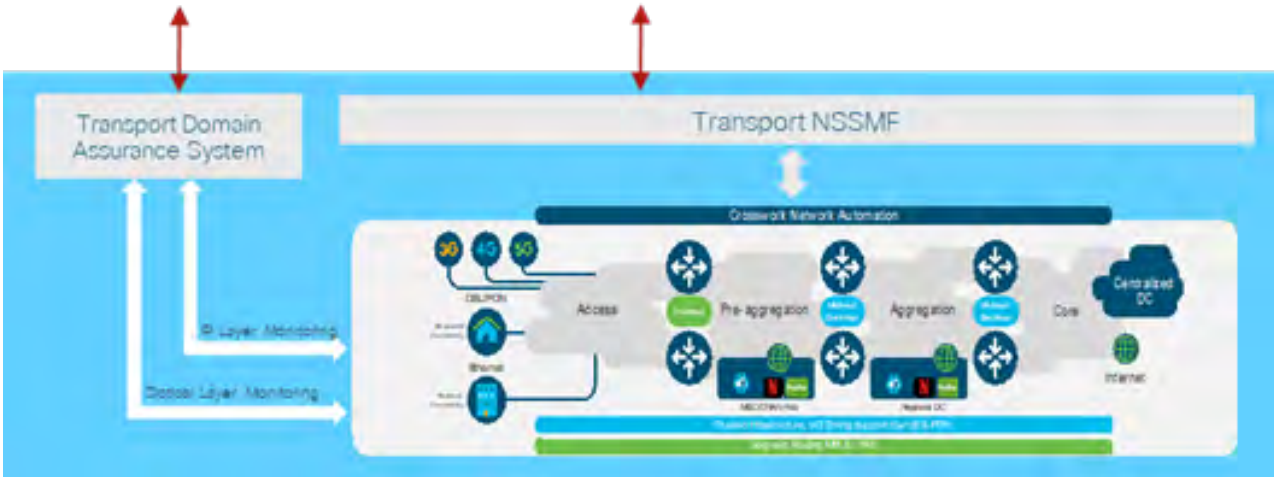
6.2 Transport domain orchestration

Transport infrastructure plays an important role for 5G services spanning multiple network domains. The end-to-end service lifecycle should be seamless, and hence, an approach toward building a unified transport architecture is essential.

At a classical top-level view, the mobility transport network is typically comprised of xHaul (fronthaul, midhaul, and backhaul) and Core domains. Both of these network domains, consisting of routers, either virtualized or physical depending on function, will participate in network slicing. Leveraging the latest Transport SDN (T-SDN) programmability and automation capabilities on top of a unified Segment Routing (SR)-based transport, device-level configuration changes related to new slice introduction or removal can be minimized, or in certain cases completely eliminated.

Before discussing specifics around network slicing for transport, we first need to explore how 5G is impacting the xHaul architecture. With the concept of RAN split, xHaul evolves to include fronthaul, midhaul, and backhaul networks. RAN split enforces very strict requirements on the xHaul for bandwidth, latency, and high availability to support precise clock synchronization between disaggregated radio functions, additional capacity due to radio densification, and mmWave deployment. Depending on the UPF placement decision, the transport network for the backhaul (N3 interface) can be either minimal (UPF placed closer to RAN edge) or extensive (UPF placed centrally or even further for certain nonlatency enterprise application use cases) [2].

Figure 9. Transport domain automation



In addition to the three service types (eMBB, URLLC, and mMTC), network slicing must also meet the following requirements within the scope of the network transport infrastructure:

- **Transport Slice Management:** Ability to create, modify, and delete a 3GPP network slice, including any actions required on the transport layer. The slice application owner and the operator must be able to monitor the health and performance of the slice through operations, administration, and maintenance (OAM) capabilities.
- **Slice Isolation:** Each transport slice must be isolated from all other transport slices in order to meet stringent SLAs. This means that the independent slice must meet proper QoS, performance, security, operation, and reliability levels.
- **Resource Reservation:** The ability to reserve transport resources for a particular transport slice in order to meet QoS requirements.
- **Abstraction:** Capability to utilize resources required to model and build a transport infrastructure to meet the demands of a network slice.

In terms of network slicing, there are two types: hard and soft. In each case, they must meet the requirements noted above, with the difference being the level of resource sharing between the slices. A hard slice has dedicated resources, most notably for transport being bandwidth, that is not shared with other slices. Soft slicing, on the other hand, is a more agile and flexible approach in which resources can be shared between slices while still maintaining SLA requirements with those resources being able to return to the network when they are no longer needed.

One of the important aspects of transport layer is that it is carrying the traffic connecting all other domains for all customers, which requires:

1. Continuous optimization of resources

   Traffic optimization and management require different levels of network automation depending on the applicable use case.

   Offline planning: Offers a complete network model taking into consideration the current network status and loads, which allows the operator to simulate what-if scenarios such as adding a new network load or service or testing an outage without affecting the working network.

   Online optimization: Performed by continuously monitoring the network conditions and SLA measurements, then using AI-driven scenarios to redirect specific flows and maintain the required SLA for most important services.

   This setup has to be completely standards based and vendor agnostic to accommodate a unified domain management solution for a potentially multivendor environment within a domain.

2. Service stitching between different domains

   The transport domain is a glue that connects the RAN network to the regional 5GC DC to the central DC and even to the end service. For this requirement, mapping the overlays from different domains such as MPLS-SR to VXLAN or maintaining an E2E MPLS-SR or SRv6 topology across different domains is required.

   This mandates the need for the cross-domain orchestrator to manage the mapping between different domains.

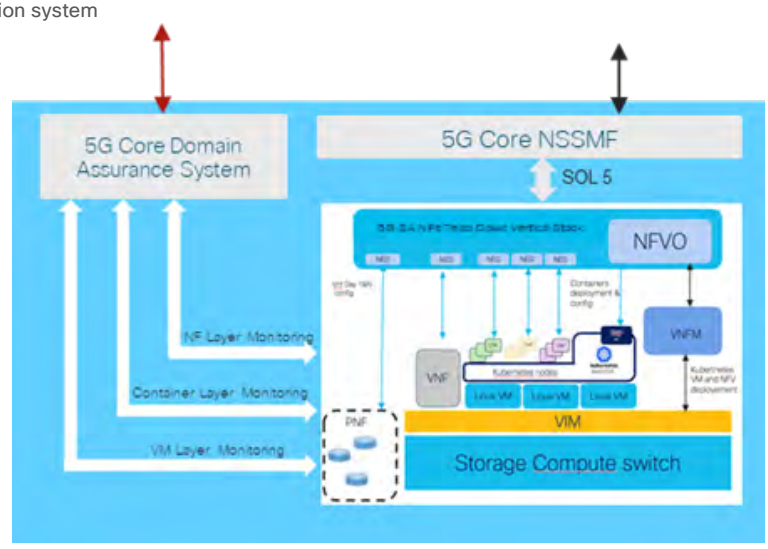
6.3 Packet core domain orchestration

As the heart of a mobile network, the packet core is expected to go through significant changes as we move from the 4G to 5G network. Standards for this transformation are in place, and 3GPP documents them in detail as of release 15.

The 5G mobile core consists of a number of new NFs, and their configuration, capacity, dynamics of creation, and removal will depend on the use case. These elements are: AMF, SMF, UPF, PCF, AUSF, UDM/UDR, NSSF, NRF, and NEF.

Most or all of these NFs will be cloud native, a collection of small, independent, and loosely coupled services, which is important because it presents a change and a challenge from both deployment and service assurance perspective compared to previous generations [2]. In 5G core automation, including some elements that may be placed at the edge location(s), a particular challenge from the infrastructure perspective presents coexistence of virtualized and cloud-native network functions, VNFs and CNFs, as they run on different platforms but have to be managed and orchestrated in a consistent manner. In the foreseeable future, it is expected to have both in the DCs, and automation solutions have to take into account instantiation, service deployments, and management of both kinds.

Figure 10. 5G core domain automation system



While Figure 5 shows a good overview of the automation architecture in general, including all layers and 5G core domain, the figure above shows a slightly different presentation of the 5G core domain alone. All of the domain automation/orchestration elements are a part of the RFS layer of the architecture, and they integrate with the CFS layer elements such as the cross-domain data collector, NSMF, and NFMF via REST, NETCONF, SNMP, or CLI. On another hand, according to ETSI [8], integration between NFVO and NSSMF (both of which are in the RFS layer) is expected to be done via the SOL005 interface. This becomes important in cases where NFVO and NSSMF are deployed as separate network functions.

ETSI MANO standards specify details of this integration, as discussed in section 3.1.3 of this document.

The key requirements for the 5GC domain orchestrator are:

- Precheck the need to instantiate new NFs to serve the new service or NSSI
- Apply the predefined NSMT (service template) to achieve the requirements of the new service
- Instantiate the required 5GC NFs through integration with existing NFVO using SOL005 interface
- Integrate with NFMF to apply NF service configuration (Day 1, 2, and N)
- Collect different telemetry from NFs and possibly from the cloud infrastructure to provide information for service assurance

Note: It is preferred to have a vendor-agnostic NSSMF and NFMF setup, which will dramatically reduce the complexity of the overall automation solution considering that the 5G core network is likely to be a multivendor environment. Having vendor-specific orchestration elements within the same domain would complicate implementation and management.

The vendor-agnostic 5G core domain orchestrator would have to support flexible workflow management, and a wide variety of configuration, fault, and performance management options.

6.4 Data center domain orchestration

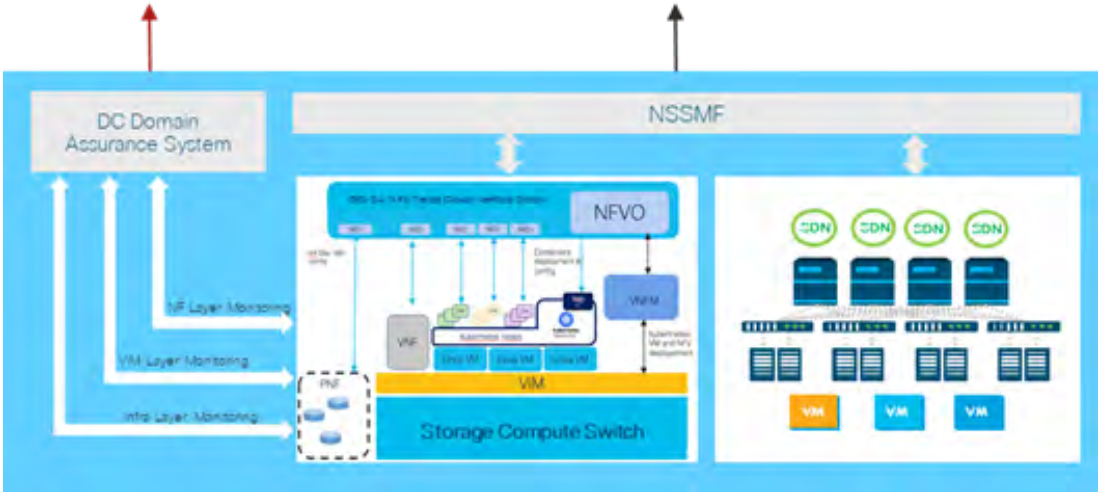
Infrastructure at various levels of DC placement (central, near-edge, far-edge, etc.) is an important aspect of 5G orchestration and automation. This includes the entire lower part of the ETSI MANO stack, taking into consideration both physical elements (compute, storage, network) and the virtualization layer, working through a VIM or cloud orchestration [2].

One key consideration that must be taken into account is how data sources and applications are distributed. Low-latency applications will require further distribution to the edge where the end users and capabilities reside. Other applications and services may also benefit by avoiding the constant backhaul or distribution of traffic from centralized data centers. The introduction of edge computing, also known as multiaccess edge computing (MEC), which is based on a distributed data center architecture, not only improves the way applications interact with end users, but also enables new applications and services that were not previously possible.

The challenges of managing a centralized data center deployment can be complex in itself, but proactively managing the infrastructure, applications, data sources, and workloads distributed across many locations takes that complexity to another level. In order to tackle this new level of complexity while enabling this new era of 5G-enabled services, it is crucial to leverage an automation and orchestration solution covering the following aspects:

- Network fabric management, including service chaining functions and resource allocation per network slice
- Network service lifecycle management, including instantiation, scaling, updating, and termination of the service
- VNF lifecycle management, including instantiation, scaling, updating, and termination of the VNF
- NFVI resources supporting virtualized and partially virtualized network functions through abstracted services for network, compute, and storage

Figure 11. Data center domain automation



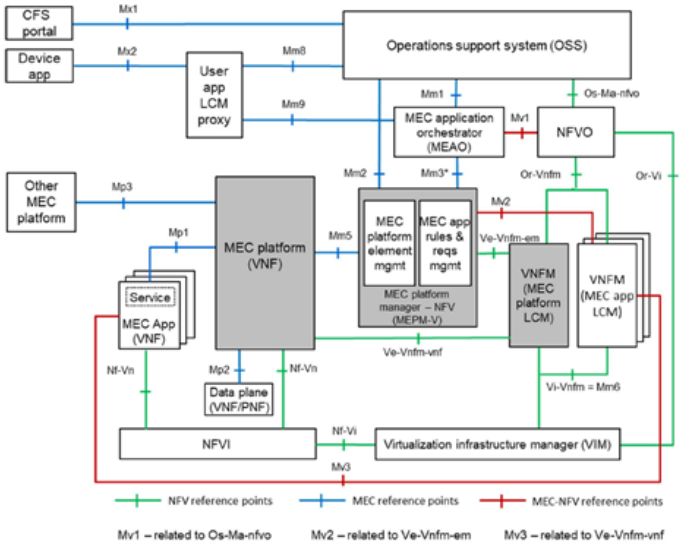
6.5 Application domain orchestration

Application Function (AF) is a part of a 5G architecture, but considering its deployment nature, a variety of applications consumed by subscribers, and MEC-specific standards for onboarding and orchestration, it has to be considered separately.

This is the area that varies the most by the use case. Application servers are expected to be installed and integrated with the network prior to the network slicing automation. Their slice-specific configuration, SLAs and KPIs would be a part of the automated slice deployment and operations [2]

Applications can commonly be placed in a centralized DC, together with the packet core control plane and other vital functions (policy, subscriber management, charging/billing, etc.), in the edge DC or the cloud. Examples of centralized DC placement applications include IP Multimedia Subsystem (IMS) or Rich Communication Services (RCS) systems. They can tolerate reasonable delay and can serve the entire subscriber population from a redundant centralized location. Many other applications that are delay sensitive or bandwidth demanding would have to be served from an edge DC closer to the subscriber. These applications can include video caching, high-end gaming, autonomous driving, etc. Orchestrating these applications can be a challenge and has to be done in sync with other infrastructure elements' orchestration. The following diagram shows an ETSI view of MEC orchestration, applicable to the Application domain [21].

Figure 12. MEC system reference architecture

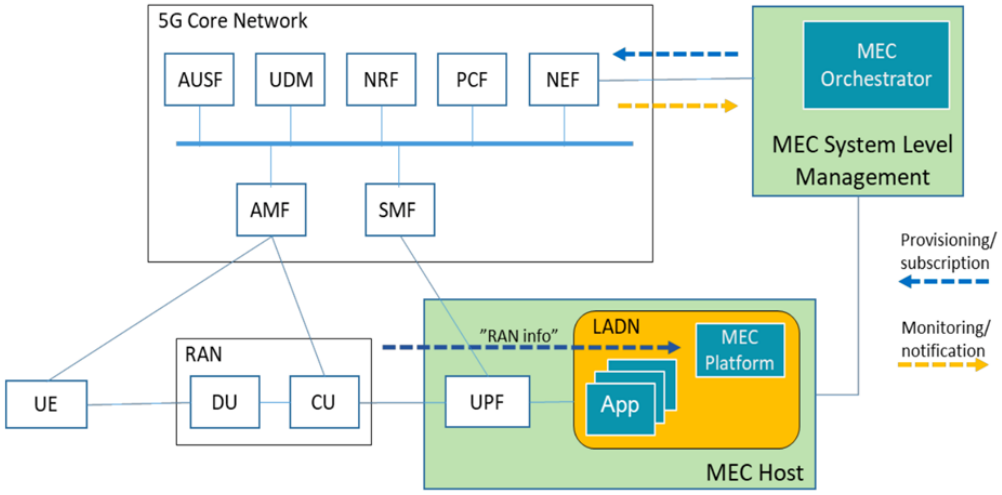


This orchestration system follows the original ETSI MANO stack, but adds several expansions requiring development and integration. The challenge is that the same MANO system is required to support very diverse applications, and the platform has to support the infrastructure elements. Furthermore, the onboarding process for these diverse applications is expected to be uniform, which presents an additional challenge.

The primary purpose of the MEC platform is to provide a standard API to access critical network information, such as latency and UE posture, and make that available to the application. Platform and orchestration systems have to support virtualized and cloud-native network functions, and in some cases even some legacy physical functions, especially in large centralized locations.

In a 5G environment, as shown in the diagram below, the MEC orchestrator and MEC host have to interact with the 5G core and 5G RAN networks for 5G capability exposure purposes, so that provisioning/subscription and monitoring/notification information can be exchanged as expected [22]

Figure 13. 5G MEC capabilities exposure



It is important to note that, at this time, ETSI standards related to MEC systems have not been widely adopted. In many cases, MEC-specific automation elements (such as MEC platform manager, MEAO, MEC VNFM) are seen as added complexity that is not always justified. It may be possible to accomplish a similar goal by modifying the existing orchestration system with some MEC-specific roles.

6.5.1 On-premises

One of the possibilities for the application placement is on the customer-premises location. Use cases for this scenario are factory automation, medicine, and corporate campuses. Private radio, licensed or unlicensed, is a significant part of the operators' opportunity in this field. Orchestration in this case has additional challenges due to physical distance from the operator's DC, and the ownership demarcation point.

6.5.2 Cloud

Public cloud providers play an increasingly important role with both SPs and enterprises. While enterprise have been moving to the cloud for many years, SPs started moving to the cloud as well, especially as they deploy edge locations and start serving their applications from the cloud. At the same time, cloud providers are entering edge business (Amazon Outpost, Google Edge TPU, Azure). Orchestration challenges in either scenario remain, as listed above.

It is important to note that edge for telco (aggregation sites/hardened COs/vRAN hosting edge sites) often means something different to edge for cloud providers (nearest tier 4 DC facility to deploy content cache/POP/reduced service offerings).

Service assurance

The 5G network platform, with new technologies in the radio, in the cloud core, and in the software-defined transport layer, is creating a more dynamic networking environment that requires automated, closed-loop network optimization to assure SLAs on a service-by-service, network-slice-by-network-slice basis.

5G service assurance is expected to ensure the QoS, QoE, and SLA across hybrid physical, virtual, and cloud-native networks, and services that cross RAN, Transport, Packet Core, DC, and Application domains. Data-driven analytics are essential to allow network slicing to be dynamically requested, scaled, and terminated. Network slicing offers different customers differentiated experiences with specific QoS, QoE, and SLA, managed on a per-slice basis. Correlating fault, performance KPIs, and service quality experience Key Quality Indicators (KQIs) with underlying resources is very important to ensure per-slice SLA and subscriber experience with balanced resource consumption. KQI formulation is to determine KPIs composing KQIs and allocating the KPIs' weight [26]

Due to large volumes of data and the dynamically distributed nature of the network, AI and automation are essential to enable 5G network slices be dynamically created, scaled, and terminated.

The figure below shows functional requirements of service assurance in 5G.

Figure 14. 5G service assurance key aspects

7.1 Cross-domain SA

Cross-domain service assurance, or CFS-SA because it resides in the CFS layer, can be viewed as the command-and-control center. It offers real-time visualization and insights from 5G network operation and service quality to customer experience. It enables intelligent assurance by applying advanced analytics, including machine learning technologies, to achieve closed-loop automation.

The cross-domain SA interface should be configurable and dynamic in nature, allowing users to configure and customize both real-time service assurance and trending analysis views dynamically:

- Define alarm generation and correlation rules
- Define KPI, KQI, SLA, and correlation rules
- Aggregate and group alarms, KPIs, KQIs, SLA per network domain, per service type, and per slicing
- User-centric customer experience measurement

On top of that, CFS-SA should provide intelligent policy-based automation capabilities in which analytics results can be evaluated by policies, which trigger service optimization or remediation automation. These analytics must include proactive anomaly detection and capacity forecasting, which requires the use of predictive machine learning algorithms. The intelligent automation interface shall:

- Provide ML functions that can be applied to KPI, KQI, and fault analysis and correlation
- Discover, create, and maintain a network orchestration, automation, and optimization workflow and actions catalog
- Evaluate analytics results against policies, to intelligently trigger actions from the catalog and achieve closed-loop effects in the network

3GPP TR 23.791 (study of enablers for network automation for 5G) has currently listed the following formula-based/AI-ML analytics use cases for 5G using NWDAF [17]:

- Load-level computation and prediction for a network slice instance
- Service experience computation and prediction for an application/UE group
- Load analytics information and prediction for a specific NF
- Network load performance computation and future load prediction
- UE expected behavior prediction
- UE abnormal behavior/anomaly detection
- UE mobility-related information and prediction
- UE communication pattern prediction
- Congestion information—current and predicted for a specific location
- Quality-of-service (QoS) sustainability, which involves reporting and predicting QoS change

7.2 Domain-level SA

Domain-level service assurance, or RFS-SA because it resides in the RFS layer, has two main functions. One is to dynamically onboard new data sources from the MEC, xHaul, core, data center, cloud, and application. All onboarded data sources will be enriched, transformed, deduplicated, and transported to a central cross-domain service assurance function for end-to-end correlation, analysis, and optimization. Second is to provide service testing and assurance as part of the CI/CD network deployment and expansion of the DevOps model.

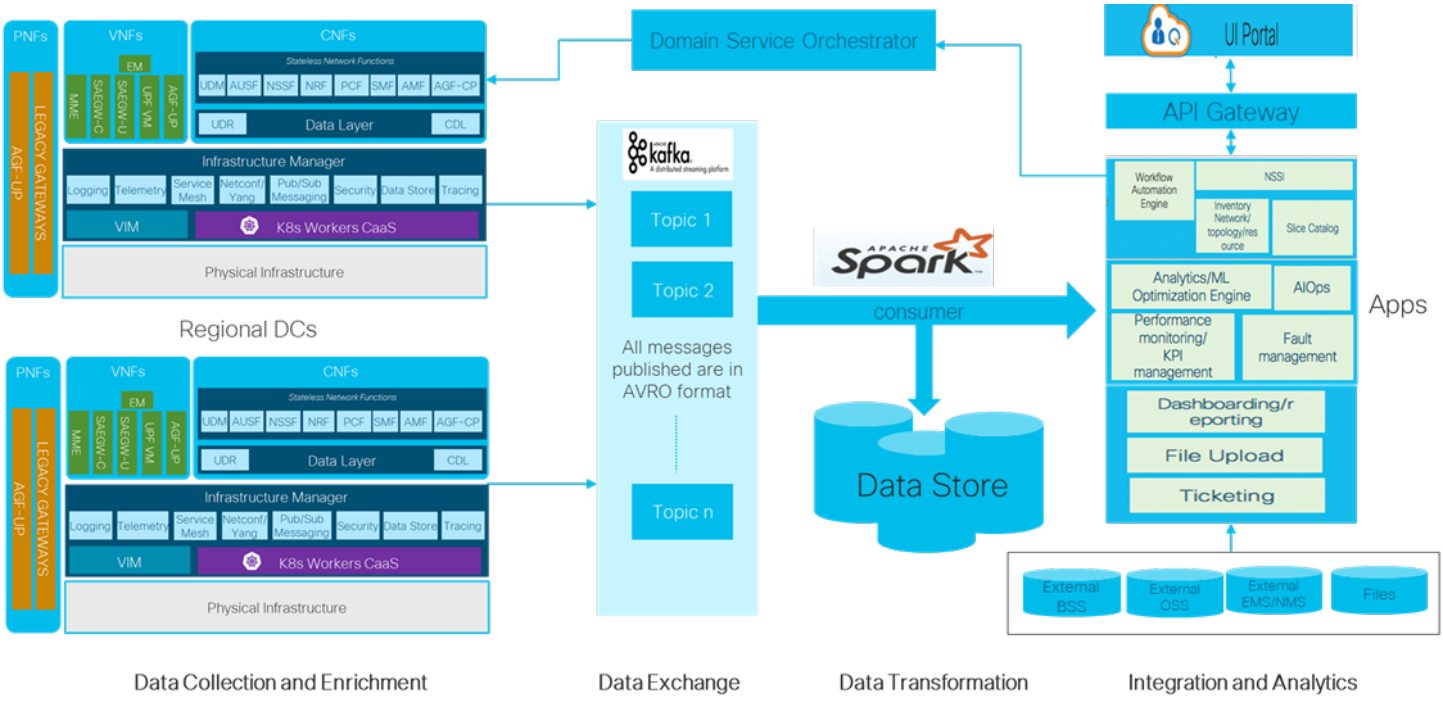
Domain-level service assurance can be viewed as part of the 5G distributed edge computing. The RFS-SA will be responsible for collecting different data sources from the NFVIs/NVFs, radio, RAN, IoT endpoints, Transport, Packet Core, DC, and Application. Enriching data by an RFS service assurance collector facilitates service chain root-cause analysis and correlation. The collector should also perform noise reduction and deduplication to reduce the amount of data sent to the central service assurance Kafka bus. The Kafka streaming system, as the data exchange platform, offers the real-time data pipeline and massive parallel processing that supports 5G service assurance use cases requiring ultra-low latency and near real-time actions. Another challenge of RFS-level service assurance is to onboard massive data sources automatically whenever new infrastructure, new end points, and new services are expanded. The network orchestrator provides a central point of configuration management for the entire network with its model-driven orchestration.

Service assurance data onboarding should be integrated with the network orchestrator and should leverage the change in network as well as the service model configuration change to automatically trigger the data source collections that are required for service assurance functions. For example, the orchestrator's ConfD northbound NETCONF interface, along with the associated YANG model, can provide a simple path to integrate with service assurance. The change can then trigger an automatic data source onboarding to the service assurance system.

7.3 Deployment workflow with closed loop

The 5G slice deployment workflow is shown in the diagram below.

Figure 15. 5G slice deployment architecture



Deployment can be looked at in two different views, focusing on provisioning and assurance. These two views, although seemingly distinct, have a significant overlap and cannot be viewed independently. With the closed-loop implementation, the systems interact in synergy.

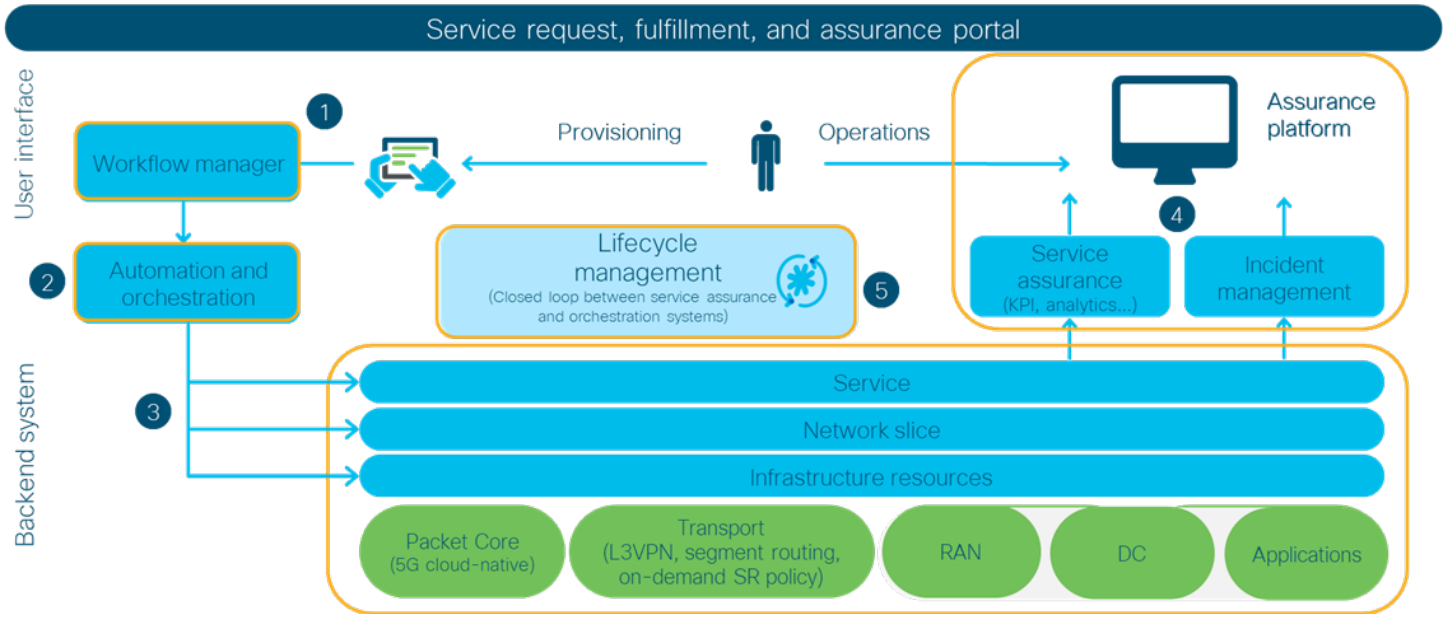
The service assurance deployment architecture consists of four verticals:

- Integration and Analytics
- Data Transformation
- Data Exchange
- Data Collection and Enrichment

Data Collection and Enrichment are done at the RFS-SA or domain-level SA layer as described in 7.2. The main purpose is onboarding of data sources.

Data Exchange, Data Transformation, and Integration and Analytics are part of the central cross-domain SA or CFS-SA functions, which provides a single pane of glass via the API gateway, for a central cross-domain, end-to-end operational view, as well as the control-and-command center for closed-loop service optimization as illustrated by the flow diagram below.

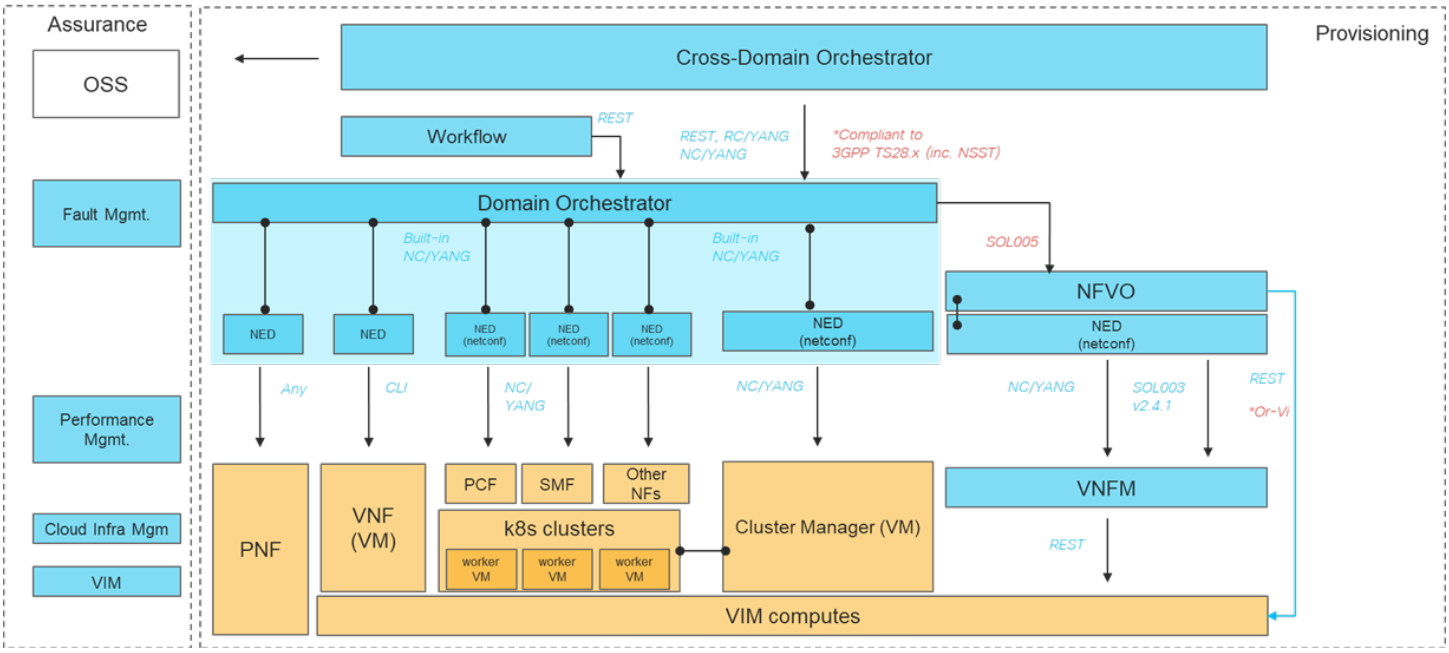
Figure 16. Network slicing—closed-loop flow



7.3.1 Provisioning focus

The provisioning focus of a deployment architecture is shown in the diagram below.

Figure 17. 5G orchestration—provisioning focus



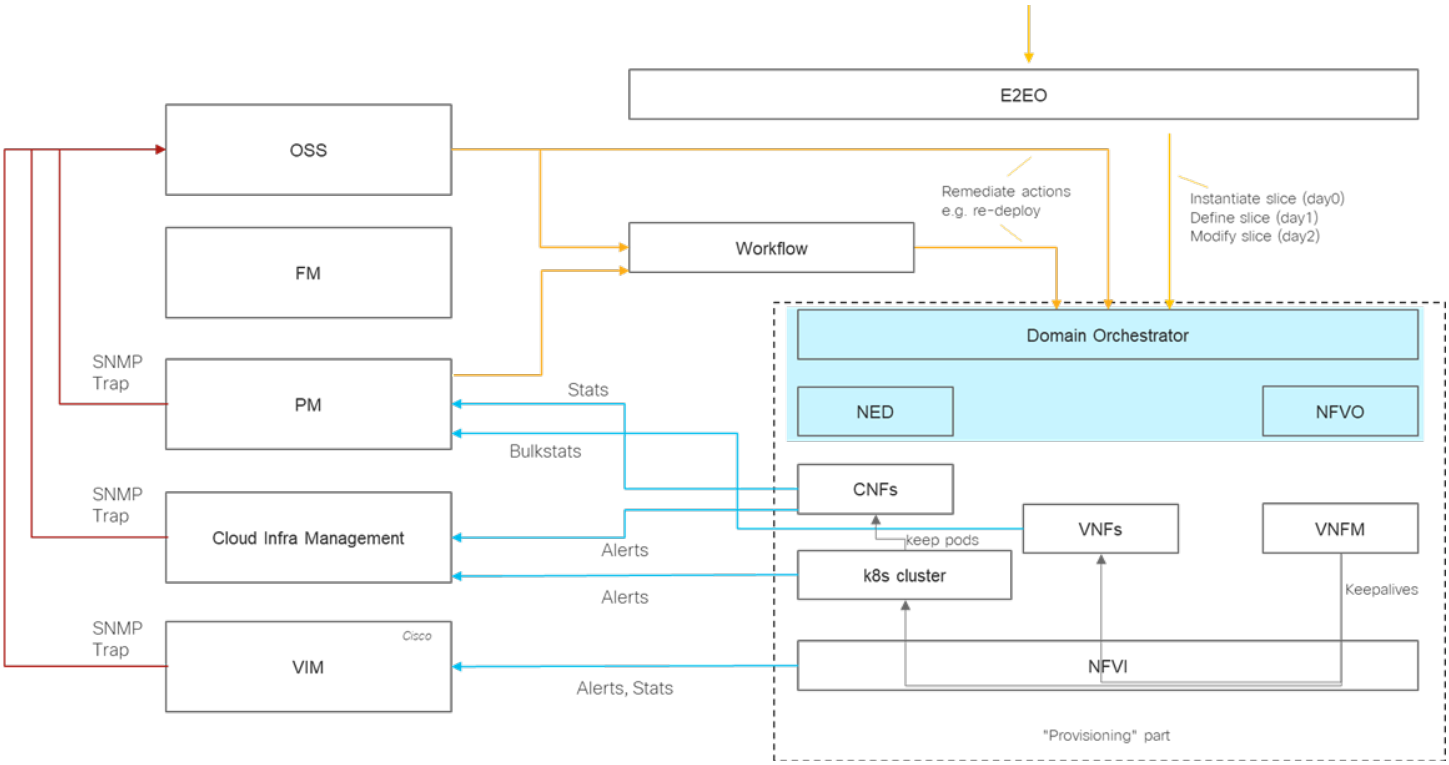
In the provisioning aspect, the OSS and cross-domain orchestrator interact via APIs. Once the order is received, the cross-domain orchestrator works on its execution through domain orchestrators as described earlier. Details of the interactions can be seen in the diagram above.

7.3.2 Service assurance focus

Service assurance systems are integrated across the OSS/BSS in order to track service performance and ensure that customer service level agreements (SLAs) are met. Service assurance systems may also proactively identify network failures and initiate resolution action.

The service assurance focus of a deployment architecture is shown in the diagram below.

Figure 18. 5G orchestration—service assurance focus



The left side of the diagram shows more details about components of the assurance system interactions with the OSS, and how the OSS ties back to the provisioning system through a cross-domain orchestrator.

Operational challenges

While enabling network slicing use cases is the end goal, there is a plethora of underlying per domain/infrastructure automation use cases that can lead to a successful solid domain primitive/construct being available to the network slice manager. Having a solid infrastructure automated deployment and automation model is required for the network slicing (such as any cross-domain use case) to be successfully instantiated, operated, and managed via the SA system.

Some items to take into consideration are:

1. Domain infrastructure deployment automation
   a. Important especially for MEC and vRAN that can explode to 1000's of data center pods
      - Consistent automated deployment of data center VIM/NFVI is key for VNF/CNF deployment
   b. Baseline configuration and operations (upgrades, expansion, etc.)
   c. Shared per domain infrastructure/applications
2. Automation tools deployment
3. IPAM and VLAN/ID management
   a. Domain infrastructure
   b. Dynamically assigned for VNFs, L3VPN, CNFs, etc. Deployed dynamically on top of slices
4. Per domain primitives/constructs that are available for use cases
   a. Dynamic per service; consumable via API in a consistent fashion
5. Multiple operational groups
   a. Infrastructure deployment and operations
   b. Applications—core/shared
   c. Applications/network functions—virtual/dynamic
   d. Network slicing and operations of slices
6. Great deal of flexibility is needed for creating service templates to respond to the ever-evolving business and service needs
   a. Development and testing of new use-case models
   b. Development of workflows and policies to accommodate the use-case models
   c. Continuous modification and cloning of existing templates

One of the challenges is the wide variety of services/slice templates that can be created, since this is a part that is not well covered and is driven by enterprise requirements that are unknown today. CI/CD plays a crucial role in developing and testing new service/slice templates to be available quickly to the enterprises.

8.1 Importance of CI/CD and DevOps

CI/CD encompasses every phase from initial software development through deployment in order to meet DevOps. CI/CD will be hierarchical and should include:

1. Automated testing
2. Artifact release management for consumption by the next phase
3. Integration with the workflow engine for UAT/production deployment

The same automation tools should be used for UAT as for production.

Artifacts may include:

1. OS versions for infrastructure devices
2. VNF image versions, VNFD
3. CNF helm charts, container images
4. Device configuration templates and models
5. Domain-specific RFS/CFS services
6. Workflows

Conclusion

Automation and orchestration of a 5G network is a complex task that has to be properly planned and implemented from the very beginning of a network design.

The complexity of 5G networks demands automation and orchestration to simplify tasks and minimize the probability of error during planning, implementation, and operation.

Automation architecture, from a functional perspective, has been defined by several standard bodies. These functions should be covered and support for them provided regardless of the implementation specifics. Different vendors will have an option to combine functions within their products as they find it appropriate; however, the functions, as shown, have to be provided to ensure proper work of the automation and orchestration system.

Automation and orchestration architecture is defined in two distinct layers: CFS (Customer-Facing Service) and RFS (Resource-Facing Service). Integration methods, protocols, and interfaces between these layers, and also between RFS and network domains as well as the CFS, portal, and OSS/BSS, have been defined by the standards and the overall automation architecture.

Automation in 5G does not stop at the network planning level. It is an integral part of the entire network cycle, including infrastructure deployment, operations and service assurance with closed-loop remediation, and DevOps with integrated CI/CD.

References and credits

References

[1] Five Steps to 5G Deployment, Cisco white paper, March 2019.

[2] 5G Automation Functional Requirements, Cisco CX internal document, November 2019.

[3] 3GPP TR 28.801 v15.1.0, January 2018.

[4] 3GPP TS 28.530 v15.0.0, October 2018.

[5] 3GPP TS 28.531 v15.0.0, October 2018.

[6] 3GPP TS 28.532 v15.0.1, October 2018.

[7] 3GPP TS 28.533 v15.0.0, October 2018.

[8] ETSI GS NFV-SOL 005 v2.7.1, January 2020.

[9] TMF640 Service Activation and Configuration API REST Specification R18.5.1, April 2019.

[10] TMF641 Service Ordering API REST Specification R18.5.1, April 2019.

[11] TMF638 Service Inventory API REST Specification R18.5.1, April 2019.

[12] TMF633 Service Catalog API REST Specification R18.5.1, April 2019.

[13] ETSI GS ZSM 002 v1.1.1, August 2019.

[14] https://www.onap.org/architecture

[15] https://wiki.onap.org/display/DW/Use+Case+Description+and+Blueprint

[16] https://wiki.onap.org/display/DW/Proposed+Functions+for+R6+and+Impacted+Modules

[17] https://inform.tmforum.org/insights/2020/06/nwdaf-automating-the-5g-network-with-machine-learning-and-data-analytics/

[18] 5G automation: The whys the how and the when, RCR wireless, March 2020.

[19] 5G Build Challenges: Other Concurrent Technology Transitions in the 5G Era, February 2019.

[20] O-RAN Specifications, https://www.o-ran.org/specifications

[21] ETSI GS MEC 003 v2.1.1, January 2019.

[22] MEC in 5G networks, ETSI white paper no. 28, June 2018.

[23] TMF639 Resource Inventory API v4.0.1, July 2020.

[24] 5G Network Evolution with AWS, July 2020.

[25] AWS Wavelength, https://aws.amazon.com/wavelength/

[26] Service Features Based KQI Definition Method, IEEE, January 2014, https://ieeexplore.ieee.org/document/6710004

[27] ITU R-REC M.2083: IMT Vision–"Framework and overall objectives of the future development of IMT for 2020 and beyond."

Credits

The authors would like to express their gratitude to the following colleagues for their significant contributions in developing the technical content of this white paper, for their work on the topics of orchestration, automation, and 5G in general, and for their reviews and constructive feedback:

- Laurent Desaunay
- John Mullooly
- Mark Swanborough
- Steve Iatrou

Glossary of terms

AF	Application Function
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
AUSF	Authentication Server Function
BBU	Baseband Unit
CFS	Customer-Facing Service
CI/CD	Continuous Integration/Continuous Delivery
CN	Cloud-Native
C-RAN	Cloud Radio Access Network
CU	Centralized Unit
CUPS	Control and User Plane Separation
DCI	Data Center Interconnect
DO	Domain Orchestrator
DU	Distributed Unit
E2EO	End-to-End Orchestrator
EPC	Evolved Packet Core
EVPN	Ethernet Virtual Private Network
IMS	IP Multimedia Subsystem
KPI	Key Performance Indicator
KQI	Key Quality Indicator
MEC	Multiaccess Edge Computing
ML	Machine Learning
mMTC	Massive Machine-Type Communications
NEF	Network Exposure Function
NFVI	Network Function Virtualization Infrastructure
NFVO	Network Function Virtualization Orchestrator
NRF	Network Repository Function
NSA	Non-Standalone
NSMF	Network Slice Management Function
NSSMF	Network Slice Subnet Management Function

- NWDAF Network Data Analytics Function
- PCF Policy Control Function
- QoE Quality of Experience
- QoS Quality of Service
- RCS Rich Communication Service
- RFS Resource-Facing Service
- (R)RU (Remote) Radio Unit
- SA Standalone
- SDN Software-Defined Networking
- SLA Service Level Agreement
- SMF Session Management Function
- SP Service Provider
- SR Segment Routing
- UAT User Acceptance Test
- UDM Unified Data Management
- UDR User Data Repository
- URLLC Ultra-Reliable Low-Latency Communications
- UPF User Plane Function
- (V)NF (Virtualized) Network Function
- VNFM VNF Manager