OpenNebula Telco Cloud Model and Architecture

Version 1.0 – April 2025

Abstract

Across multiple projects, OpenNebula has proven its capability to support both open source and vendor-proprietary solutions for deploying and integrating end-to-end 5G networks, from UE simulators to 5G Core Network elements and other workloads such as AI/ML technologies.

As part of several innovation efforts, OpenNebula has enhanced its Telco Cloud Model and Architecture, Artificial Intelligence techniques and Zero-Touch resource management methods for the efficient deployment and operation of distributed cloud-edge-continuums. OpenNebula is helping its users to combine centralized clouds with edge resources and to choose the right combination of geographically-distributed cloud-edge locations to efficiently execute their workloads, meet their enterprise needs, and avoid vendor lock-in.

This white paper describes OpenNebula's Application Model, its network model for integration with 5G Advanced functionality, and the models' main management challenges; as well as OpenNebula's Framework Architecture, its main building components, and its optimized edge clusters.

Contents

Glossary

4G	4th Generation Mobile Network
5G	5th Generation Mobile Network
6G	6th Generation Mobile Network
AI	Artificial Intelligence
AIaaS	AI-as-a-Service
AMF	Access and Mobility Management Function
API	Application Programming Interface
AWS	Amazon Web Services
AUSF	Authentication Server Function
B5G	Beyond 5G technology
BBU	Baseband Unit
BSF	Binding Support Function
CDN	Content Delivery Network
CI	Continuous
CU	Central Unit
CPU	Central Processing Unit
DevSecOps	Development, Security and Operations
DNS	Domain Name System
DPDK	DPDK Data Plane Development Kit
DU	Distributed Unit
Gbps	Gigabits per second
HA	High Availability
IP	Internet Protocol
IT	Information Technology
KVM	Kernel-based virtual machines
L2	Layer 2
ML	Machine Learning
MLOps	Machine Learning Operations
NFV	Network Function Virtualization
NRF	NF Repository Function
NSSF	Network Slice Selection Function
NUMA	Non-Uniform Memory Access
O-RAN	Refers to the O-RAN Alliance
OneKE	OpenNebula Kubernetes Edition
PCF	Policy and Charging Function
PCI	Peripheral Component Interconnect
PDCP	Packet Data Convergence Protocol
QEMU	Quick Emulator
QoS	Quality of Service
RAN	Radio Access Networks
RFC	Requests for Comments
RIC	RAN Intelligent Controller
RKE2	Rancher' s Kubernetes distribution
RRC	Radio Resource Control
RRH	Remote Radio Head
RU	Radio Unit
SCP	Service Communication Proxy
SEPP	Security Edge Protection Proxy
SFC	Service Function Chaining
SFN	Service Function Network
SFP	Service Function Path
SMF	Session Management Function
SR-IOV	Single Root I/O Virtualization
UDM	Unified Data Management
UDR	Unified Data Repository
UE	User Equipment
UPF	User Plane Function
VLAN	Virtual Local Area Network
VM	Virtual Machine
VNF	Virtual Network Function
VPN	Virtual Private Network
VR	Virtual Router
VXLAN	Virtual eXtensible LAN

PART A. Introduction

A.1. AI/ML in B5G Networks

5G, the 5th generation mobile network, is a new type of network designed to connect everyone and everything including machines, people, devices, businesses. 5G technology was created to deliver multi-Gbps peak data speeds, more reliability, massive network capacity, ultra low latency and a more uniform user experience across the whole network, empowering new user experiences, connecting new industries, facilitating innovation and improving efficiency.

With the adoption of cloud, edge, 5G and Open RAN technologies, business network infrastructure and design was reimagined. Cloud and 5G have transformed connectivity, enabling data processing at lower latencies, real-time communication, faster data transfers and scalable services. Benefiting from a wider frequency spectrum, 5G can handle a larger number of connected devices; its design ensures more reliable and stable connections even in dense urban environments. The deployment of this generation of mobile networks and the adoption of the complementary Open RAN paradigm are speeding up the adoption of the cloud. 5G and Open RAN reach their full potential when they follow cloud-based approaches in their deployments.

5G and edge computing are symbiotic technologies which enable new use cases: they both aim to significantly enhance networks' operation, performance and capabilities for transferring huge amounts of data that needs to be processed in real-time. While 5G brings much higher speeds than 4G, edge computing reduces the latency and the bandwidth consumption at the level of core networks. Edge computing can be perceived as a complementary technology to 5G, allowing it to carry large volumes of data with reliability, speed and efficiency.

Open RAN is a new paradigm in the design and deployment of radio access networks. Its goal is to disaggregate RAN functionalities and to open up the interfaces between them,enabling new types of deployment, greater flexibility and higher level of innovation; and creating new business opportunities for telecom actors. Open RAN can be divided into three parts:

- Cloudification/virtualization of the RAN components
- Management and orchestration
- Open interfaces

Specific proprietary RAN functions such as remote radio head (RRH) and baseband units (BBUs) can now be disaggregated into three different components: centralized units (CU), distributed units (DU), and radio units (RU). Open RAN allows service providers to benefit from better market options, avoiding vendor lock-in and encouraging vendor diversity.

The relationship between 5G, edge computing, Open RAN, public and private clouds and other enabling technologies facilitates seamless connectivity across multiple types of cloud (e.g. hybrid, multi-cloud) and from public clouds to edge devices and everything in between.

B5G networks have at their core the concepts of virtualization and cloud-nativeness — concepts applied to all network elements across the cloud-edge continuum. By adopting virtualization and cloud-native principles these networks gain flexibility, scalability, and efficiency. Ensuring flexibility in the deployment of cloud-native solutions over a heterogeneous cloud-edge continuum infrastructure, and supporting distributed execution and multi-tenancy, call for a unified design of service orchestration, management, and control plane.

AI/ML appear as another complementary paradigm to the above-mentioned technologies. Already

becoming obvious in several network processes, the integration of AI and ML mechanisms has enabled automation in service and resource management. Maintaining this trend and continuing their evolution, AI/ML technologies have the capability to powerfully impact the way future networks operate, by enabling predictive and autonomous functionalities at various network layers. AI is already considered as one of the main building blocks of the future 6G networks and, in consequence, the industry is prioritizing research for AI-based technological enablers.

The integration of AI/ML will have a huge effect on how the 6G system architecture will be designed and implemented. Aspects related to the APIs, data storage requirements, workload placement, compute requirements, MLOps, privacy and security will need to be carefully taken into account and designed, while considering use case requirements. While 6G is poised to bring into reality the support for a multitude of parallel and heterogeneous slices, the challenges related to scalability and sustainability will need to be addressed by deploying AI-driven Zero-touch management and orchestration frameworks. The B5G networks will need to include mechanisms at multiple levels to optimize resource allocation and workload placement across cloud-edge continuums.

6G networks will be AI-native. From this perspective, their architecture will be characterized by several key features:

- Intelligence everywhere
- Zero-touch management
- Distributed data infrastructure
- AI-as-a-Service (AIaaS)

Intelligence will be integrated across the network, from central nodes to edge devices, allowing AI/ML workloads to process information wherever it's most effective. This will require a distributed data infrastructure to ensure data availability and seamless processing. Zero-Touch Management will continue to enable autonomous network operations. AIaaS will expose AI/ML models, datasets and tools as services, enabling service providers and users to leverage AI capabilities and to create new use cases.

A.2. OpenNebula

OpenNebula[1] is a **simple, but powerful, open source solution to build and manage Enterprise Clouds and Edge environments.** It combines virtualization and container technologies with multi-tenancy, automatic provision, and elasticity to offer on-demand applications and services.

OpenNebula provides a single, feature-rich and flexible platform with **unified management of IT infrastructure and applications** that **avoids vendor lock-in** and **reduces complexity, resource consumption, and operational costs**. OpenNebula manages:

- **Any Application**: Combine containerized applications from Kubernetes with Virtual Machine workloads in a common shared environment to offer the best of both worlds: mature virtualization technology and orchestration of application containers.
- **Any Infrastructure**: Open cloud architecture to orchestrate compute, storage, and networking driven by software.
- **Any Cloud**: Unlock the power of a true hybrid, edge and multi-cloud platform by combining your private cloud with infrastructure resources from third-party virtual and bare-metal cloud providers such as AWS and Equinix Metal, and manage all cloud operations under a

[1] https://support.opennebula.pro/hc/en-us/articles/360036935791-OpenNebula-Overview-Datasheet

single control panel and interoperable layer.
- **Any Time**: Add and remove new clusters automatically in order to meet peaks in demand, or to implement fault tolerance strategies or latency requirements.

OpenNebula provides the necessary tools for running containerized applications from Kubernetes while ensuring enterprise requirements for your DevSecOps practices. It helps organizations to easily embrace Hybrid and Edge Computing, allowing them to grow their Enterprise Cloud on demand with infrastructure resources from third-party Public Cloud and bare-metal providers such as AWS and Equinix Metal. This disaggregated cloud approach allows for a seamless transition from centralized private clouds to distributed edge-like cloud environments. Companies can grow their private cloud with resources at cloud and edge data center locations, to meet peaks in demand or the latency and bandwidth needs of their workload. This approach involves a single management layer where organizations can continue using existing OpenNebula images and templates, keep complete control over their infrastructure, and avoid vendor lock-in.

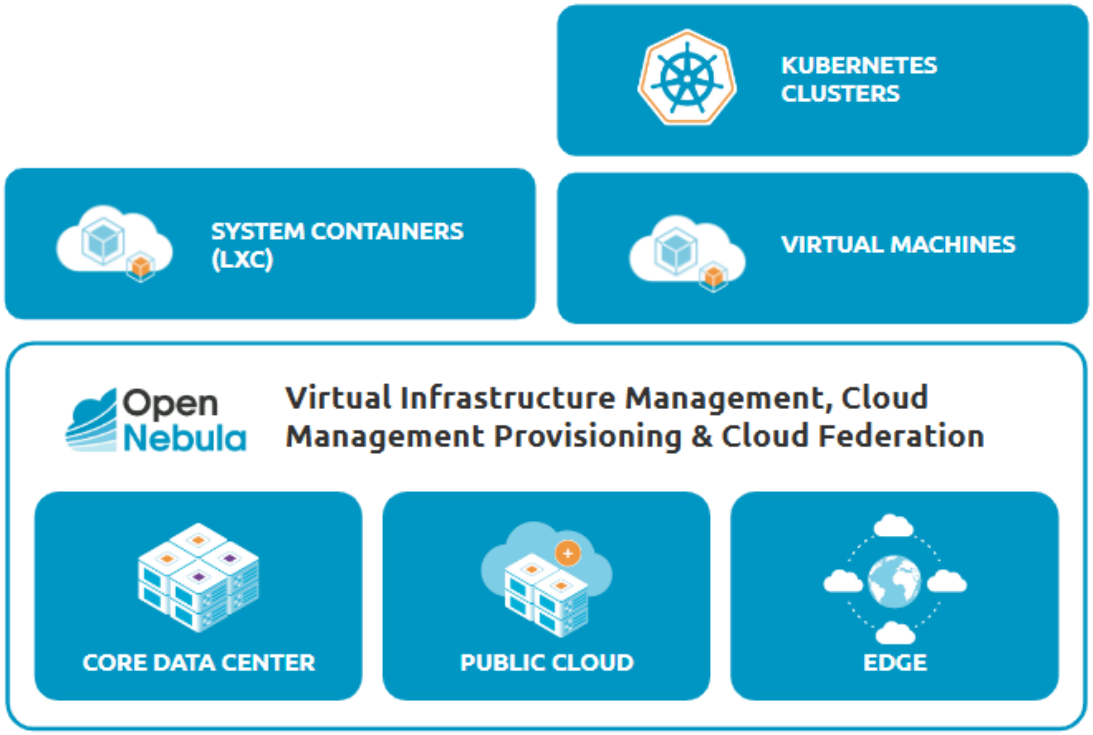


Figure 1. OpenNebula's flexibility.

The development of OpenNebula follows a bottom-up approach driven by the real needs of sysadmins, DevOps, and corporate users. OpenNebula is an **open source product** with a healthy and active community, commercially supported by OpenNebula Systems through its OpenNebula Subscription program. New versions are released on a regular basis and delivered as a single package with a smooth migration path. OpenNebula defines its short-term roadmap and plans the features for the next release guided by demands of its Sponsors, Customers, Users and Partners. A detailed list of planned features for the upcoming release of OpenNebula is available at the GitHub OpenNebula/One issues page. More information on the benefits of running an OpenNebula cloud can be found on the Key Features page.

Have a look at our Case Studies and Success Stories to learn more from our users about how they are putting OpenNebula to work.

A.3. OpenNebula ONEedge5G

ONEedge5G is an industrial research and innovation project led by OpenNebula Systems, focusing on the development of Artificial Intelligence (AI) techniques and Zero-Touch resource management methods to enhance the deployment and operation of distributed edge environments over 5G-Advanced infrastructures. The project aims to facilitate the efficient management of edge systems, ensuring low latency and energy optimization for advanced data processing applications.

A key objective of ONEedge5G is to enable new industrial actors to leverage 5G-Advanced infrastructure effectively. This involves implementing AI-driven orchestration for capacity planning, workload forecasting, and risk prevention in geographically distributed edge infrastructures. The project also emphasizes the integration of open-source solutions to manage private 5G deployments, promoting innovation aligned with European strategic priorities. The outcomes of this research have been validated through highly demanding use cases and are being incorporated into OpenNebula to encourage rapid adoption in the emerging market.

This White Paper includes:

- A description of OpenNebula's **ONEedge5G Application Model and** network model for integration with 5G Advanced functionality, and their main management challenges.
- A description of OpenNebula's **ONEedge5G Framework Architecture**, its main building components, and its optimized edge clusters.

One of the main aims of the framework is to enable the incorporation of the AI analytical capabilities and Zero-Touch techniques:

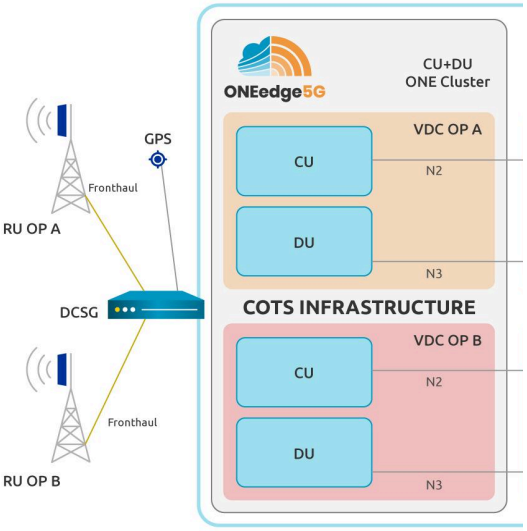
- **Combining workloads**, including AI-based, container-based, and virtual machine-based applications in a shared environment.
- **Enabling interoperability and multi-cloud portability** through a unified view of underlying resources from local data centers and different cloud providers. The ability to freely move applications allows for the implementation of dynamic planning and migration heuristics based on AI.
- **A programmatic approach to the edge-as-a-service model**, built on different distributed environments. This automation will be used to implement AI techniques.
- **Integration with a catalog of cloud and edge providers** to intelligently select the ideal provider for a specific type of workload.

Three different types of requirements have been considered:

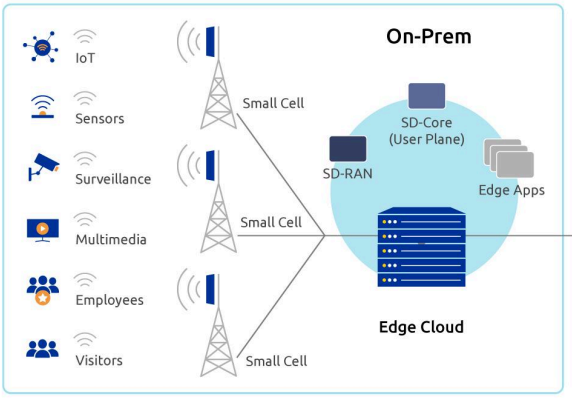
- **General Requirements**: These requirements outline the core elements necessary to align the ONEedge5G Framework with EU digital strategies, focusing on sovereignty, sustainability, interoperability, and security. This not only ensures that the framework meets EU standards, but also operates efficiently and facilitates the integration across multiple cloud and edge environments.
- **European Telco Industry Requirements**: Specific needs and objectives within the European telecommunications sector. It takes into account initiatives such as Sylva Open-Source Telco Cloud and the EU Alliance for Industrial Data, Edge and Cloud. It highlights the importance of developing a telco cloud framework that supports EU businesses and public administrations in processing sensitive data, fosters strategic autonomy, and minimizes dependency on non-EU technologies.

- **Use Case Requirements**: The use case requirements focus on the practical application and deployment of the ONEedge5G framework in real-work scenarios. They demonstrate how the framework could accelerate the digital infrastructure through 5G Advanced and 6G technologies in several use cases, including distributed edge environments for NFV, neutral 5G operator hosting, and fully disaggregated central office for Telco Edge 5G.

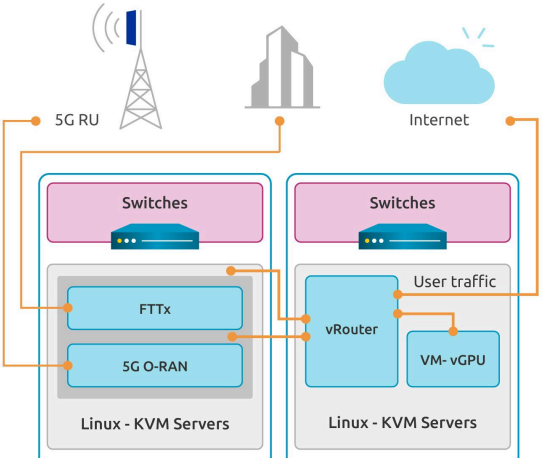
Distributed Edge Environment for NFV



Neutral 5G Operations Hosting



5G Edge Deployment



Fully Disaggregated Central Office

The **main objectives of the OpenNebula's ONEedge5G Framework** are:

- Satisfying the demand for **open European technologies for 5G Telco Cloud**.

- Integrating the functionality of **5G-Advanced to improve the edge**.
- Making use of **open solutions to manage the deployment of private 5G**.
- Expanding the **cloud-edge continuum through new 5G-Advanced infrastructures**.
- Facilitating the **distributed management of the cloud-edge continuum** for businesses and users.

These are the **main components of OpenNebula's ONEedge5G Continuum Architecture, which provides** an abstraction layer to deploy applications in the cloud-edge-5G continuum.

- The **5G-Advanced Integration** integrates advanced networking technologies for the operation of the cloud-edge.
- The **5G Edge Cluster** is optimized to operate a distributed infrastructure, and automate the deployment and operations of each segment of the infrastructure.
- The **5G Workflows Management** for the workflow applications consisting of multiple VM and networks.
- The **5G Workflows AI-Driven Orchestration** for enabling the scalable monitoring, intelligent prediction of metrics, and scheduling algorithms for placement and automatic scaling.
- The **Containerized Applications Enablement** to provide a common approach for the container application layer, by running a Kubernetes cluster in a set of VMs.

PART B. OpenNebula Conceptual Model for 5G Networks

This section presents the proposed foundational framework for modeling the applications and 5G networks. It aims to provide a detailed guide for developing efficient and robust digital infrastructures able to meet the requirements of modern edge computing and 5G technologies.

Application Model: Describes a simplified application model in which processing elements communicate directly over a Layer 2 network, focusing on virtual networks, virtual machines, and their deployment attributes. This model supports generic applications through direct interconnectivity and deployment flexibility.

5G Network Model: Outlines the essential components for 5G network deployment, including Radio Units, Distributed Units, Central Units, and the Mobile core. This section details how these elements integrate within the application model to facilitate the deployment of end-to-end 5G mobile networks, including Open RAN deployments.

Application Management Challenges: Addresses the operational challenges in application deployment and management, including optimal placement, zero-touch deployment, resource allocation and auto-scaling, and anomaly detection. It also describes the importance of efficient placement strategies, automation in deployment, dynamic resource scaling, and robust monitoring to maintain the integrity and performance of applications and networks.

B.1. Application Model

In general, a network application can be modeled as a set of interconnected processing components, each with a different set of requirements. This Service Function Chain (SFC) paradigm

is usually adopted to implement and manage network services in a shared infrastructure. In this context, it is assumed that the infrastructure provides some traffic steering capabilities to move network packets from one function to the next one (see for example the Network Service Header extensions described in RFC 8300).

In OpenNebula we use a simplified model where all processing elements are interconnected through a Layer 2 (L2) network. Each element is able to communicate with each other directly by their link addresses over private, isolated links. Figure 2 depicts these two models. The main difference between them is that the Service Function Path (SFP) is replaced by a virtual network, and hence each element needs to explicitly address the next one in the chain. This modification will allow us to address more generic applications.

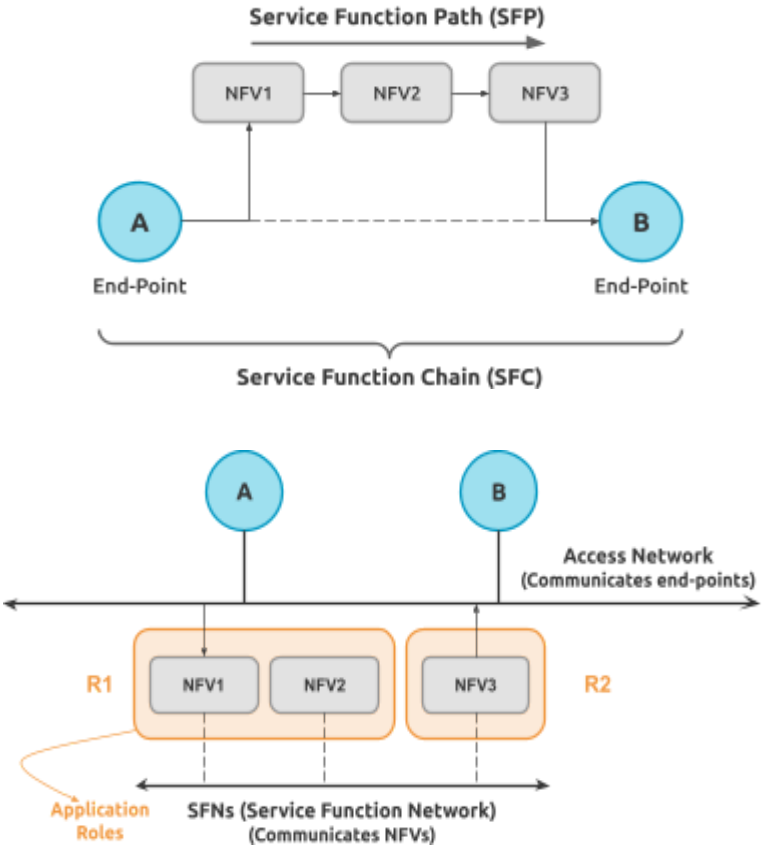


Figure 2. Service function chain (SFC) model (upper image) and OpenNebula application model (lower image).

The OpenNebula application model consists of the following elements:

- **Virtual Networks.** These represent the interconnection of all the application elements. A virtual network includes a private IP address space, a suitable isolation technology (e.g. VXLAN encapsulation or 802.1Q VLAN tagging) and optionally other configuration parameters (e.g. DNS servers or default gateway).
- **Virtual Machines.** These are the processing components. The definition of a VM includes:
    - Capacity requirements (e.g. memory or CPU).
    - Interconnection links, including QoS requirements.
    - Any additional hardware specification.

- **Deployment and Parametrization Attributes.** Cloud applications usually follow an *"install-once-deploy-many"* approach that requires pushing parameters which specialize a generic component for the target application. Moreover, some application components have a deployment order constraint (e.g. a component needs to be deployed after another one) that needs to be defined as part of the application.

Usually a given function in an application can be implemented by more than one VM that shares the same requirements and deployment constraints. This set of VMs constitutes an **application role** (see the lower image in Figure 2)**.** A role also includes:

- **Cardinality.** The number of VMs in a given role.
- **Elasticity Rules.** The cardinality in a role can be adjusted according to the performance of the application. These rules may be a combination of application-specific metrics (e.g. number of requests per second) and VM-associated metrics (e.g. CPU load).

B.2. 5G Network Model

A 5G Network may be formulated within the application framework described above. In general, a 5G network includes the following components:

- **Radio Unit (RU).** The RU is the hardware component responsible for transmitting digital data over radio frequency signals in a reliable manner. The RU performs the digital-analog conversion, transmission, and amplifications of the signals.
- **Distributed Unit (DU).** The DU is responsible for data coding and modulation, multiplexing and real-time scheduling of the radio link. The scheduling algorithm is used to slice the radio spectrum to allocate a variable capacity to different users or application classes (*network slicing).*
- **Central Unit (CU).** Performs policy-related operations for the 5G pipeline as well as converting end-to-end protocols (e.g. IP) to packets that can be transmitted over the DU-RU, including header compression, ciphering, re-ordering, and duplication. This transport service implemented with the Packet Data Convergence Protocol (PDCP) is used for both the user and control planes. The CU also implements coarse-grain policies resource control (RRC) for the transmission pipeline. This last component is usually referred to as RAN Intelligent Controller (RIC).

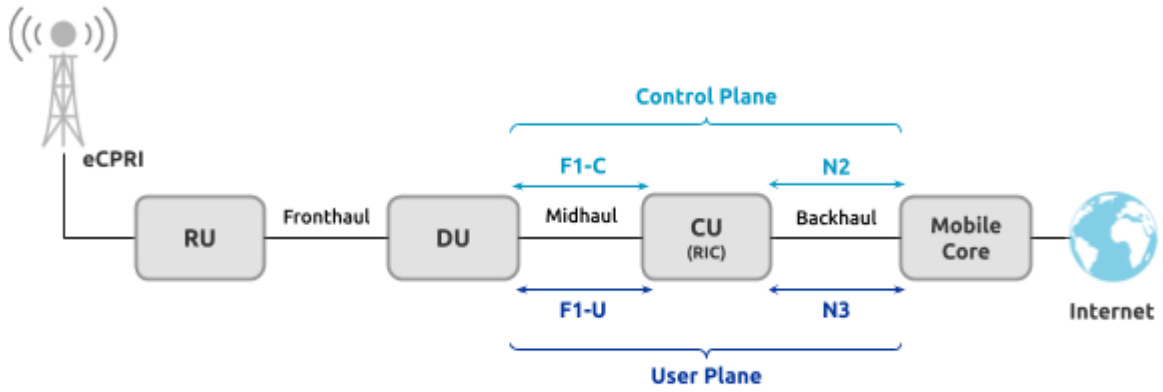


Figure 3. 5G Open RAN architecture.

- **Mobile core.** It provides multiple functions, the following being the more important ones: the forwarding of user IP traffic from the radio network to the internet (UPF); location services and access and mobility management (AMF); and user equipment session management (SMF) including IP address and QoS allocation. A complete mobile core includes other components for authentication of the user equipment, identity management, or service discovery, among others as described in the below table:

Table 1: Network Functions included in 5G SA.

Network Function	Description
NRF	NF Repository Function, manages and stores information about all network functions.
SCP	Service Communication Proxy, acts as an intermediary to manage and optimize communication between network functions.
SEPP	Security Edge Protection Proxy, provides security for inter-network communication between 5G core networks (roaming).
AMF	Access and Mobility Management Function, manages user access, mobility, and connection states (e.g. registration or handovers).
SMF	Session Management Function, controls and manages user sessions and IP address allocation.
UPF	User Plane Function, handles data traffic routing and forwarding, providing a direct link between the user and external networks.
AUSF	Authentication Server Function, manages user authentication and verifies identity for secure access to the 5G core.
UDM	Unified Data Management, centralizes user subscription data and profile management.
UDR	Unified Data Repository, stores subscriber data, policy information, and session data.

PCF	Policy and Charging Function, controls policy rules and charging for network usage, managing quality of service (QoS) and data usage policies.
NSSF	Network Slice Selection Function, selects the appropriate network slice for each device.
BSF	Binding Support Function, manages IP address bindings to ensure session continuity and efficient routing.

These components communicate through standardized interfaces and can be directly mapped into the general application model described above. Moreover, Open RAN deployments allow the distribution of real-time RAN functions to the edge, creating an opportunity for OpenNebula to become the foundation layer of end-to-end 5G networks.

B.3. Application Management Challenges

In the previous sections we have described a framework that can be used to model a general application as well as a 5G Network. Deploying such applications presents non-trivial challenges from an operational perspective, including application shaping and placement as well as sustainability aspects.

Optimal Application Placement

Once the application roles and virtual elements have been dimensioned, we need to determine the optimal placement of each element of the application. In particular the placement algorithm needs to consider:

- The overall limit of the application components should not exceed the capacity of the physical infrastructure, subject to some pre-configured over-commitment limits.
- The allocation must satisfy the QoS requirements of the application in terms of latency or bandwidth of each interconnection link.
- The placement algorithm should consider different optimization criteria depending on the application, as well as administrative constraints. It is of particular interest to include energy/power consumption metrics to minimize carbon emissions of the 5G network and applications.

Zero-touch Application Deployment

An application like the one described above, including multiple components each with specific configurations and bootstrapping data, requires a highly automated, zero-touch approach for the deployment and delivery of the VM components. In particular this requires two basic elements:

- A flexible framework for building and publishing VM components. This includes the ability to create VM disk images with an up-to-date installation of the application. Additionally, the framework should allow the injecting of specific code to perform the zero-touch configuration of the VM at runtime. Finally, to support the dynamic deployment of the applications in multiple places, a suitable way to publish and distribute these VM images should be in place.

- The workflow and VM management system needs to provide a flexible mechanism for passing parametrization data to each application component at runtime. This data is used to perform the zero-touch configuration of the overall application. It' s important to note that this data may contain sensitive information, requiring the framework to provide secure ways to store it and pass it to the VMs.

Resource Allocation and Auto-scaling of VNFs

Once an application is defined in terms of roles and interconnection networks, it should be properly dimensioned to address the application demand and so avoid under and over-provisioning. The resource allocation can be classified in terms of:

- *Vertical capacity*, which refers to the resource characteristics of each VM in terms of the capacity allocated to the memory, CPU, storage, and network subsystems.
- *Horizontal capacity,* which refers to the number of VMs in a given role (cardinality).

Additionally, the resource allocation can be static, i.e. set at deployment time, and dynamic, so the capacity adjusts to the varying demands of the application.

Anomaly Detection

In order to supervise the correct operation of the 5G network and applications, it is imperative to deploy monitoring and alerting systems as part of the management plane. Usually the data collected for each system component includes: raw infrastructure performance metrics, e.g. used memory or CPU; application metrics, specific for an application class, e.g. transactions or frames per second; and operational metrics, e.g. status of services. The data collected by the monitoring system can be used by the AI operations module to detect anomalies, attack and intrusion patterns, or faulty conditions .

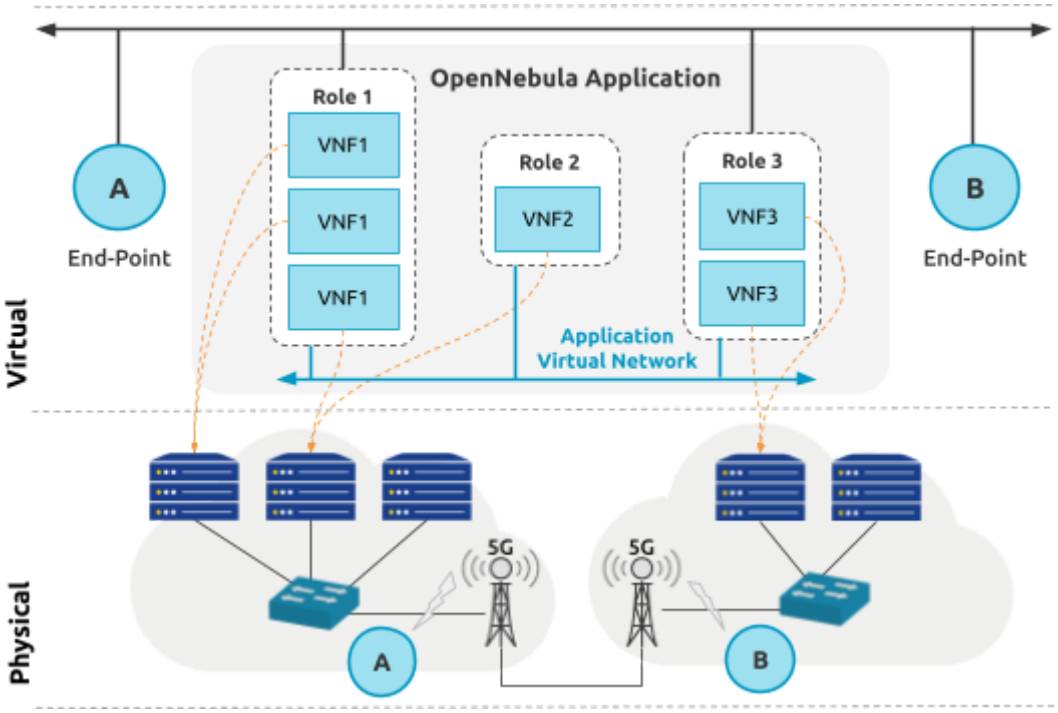


Figure 4. OpenNebula Network and Monitoring Architecture.

PART C. OpenNebula Architecture for the Cloud-Edge-5G Continuum

The OpenNebula architecture provides an abstraction layer to deploy applications in the cloud-edge continuum. The main components of OpenNebula's OpenNebula Continuum Architecture are:

- The 5G-Advanced Integration integrates advanced networking technologies for the operation of the cloud-edge.
- The 5G Edge Cluster is optimized to operate a distributed infrastructure, and automate the deployment and operations of each segment of the infrastructure.
- The 5G Workflows Management for the workflow applications consisting of multiple VMs and networks.
- The 5G Workflows AI-Driven Orchestration for enabling the scalable monitoring, intelligent prediction of metrics, and scheduling algorithms for placement and automatic scaling.
- The Containerized Applications Enablement to provide a common approach for the container application layer, by running a Kubernetes cluster in a set of VMs.

These components will be detailed in the next sections.

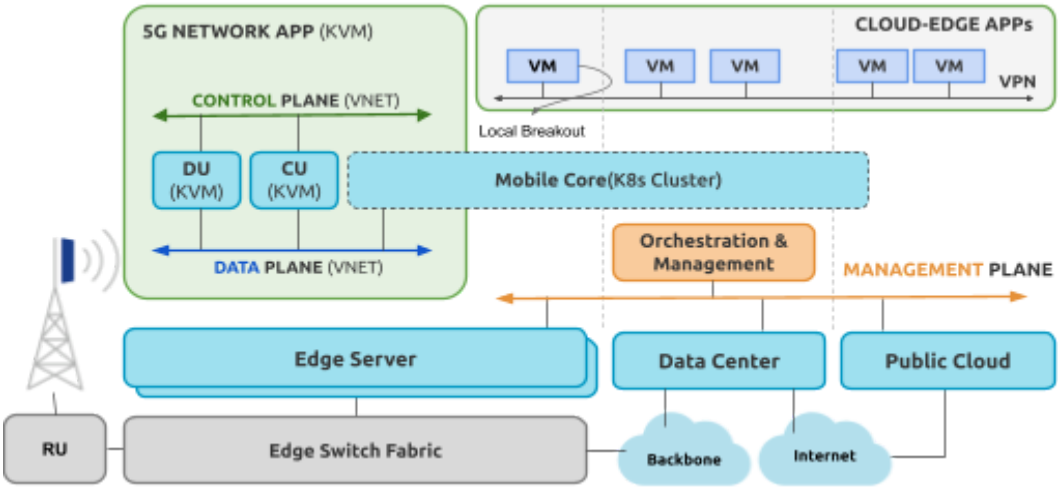


Figure 5. Overall architecture of the framework for deploying 5G networks and applications in the cloud-edge continuum.

The applications deployed on top of cloud-edge continuum are formulated as workflows with different application roles and virtual networks. Figure 5 depicts the main components of the framework. In general, we consider that the deployment of the application can span three different segments:

- **Edge Segment.** This segment comprises the facilities at each point of presence. Usually this includes the radio unit (RU) modules and a small number of edge servers. This infrastructure runs the components of the application that needs to run on the edge for latency reasons. An example is the DU component of a 5G RAN network that is usually co-allocated with the RU to perform real-time traffic scheduling.
- **Private Cloud Segment.** It denotes the private infrastructure directly connected to the edge location via a network backbone. Typically, this segment is managed by ISPs (Internet

Service Providers) and telecommunication companies. For example, some components of the 5G network application are deployed centrally to control multiple 5G networks. Application components deployed in the private cloud have non-trivial network requirements.

- **Public Cloud Segment.** It represents the facilities of public cloud companies functioning as cloud services providers. When there are no strong latency requirements, some components of the application can be deployed in the public cloud (centralized cloud). From an operational point of view, this approach offers several advantages like the HA (*High Availability*) deployment of such components.

The ability to deploy application components as defined by the OpenNebula framework across these three segments enables an effective use of the so-called Cloud-Edge continuum.

C.1. 5G - Advanced Integration

The use of the Open RAN specifications and 3GPP standards enable us to manage and deploy a 5G network as an OpenNebula application. In this framework we can assume that the RAN components are deployed as virtual machines while the mobile core is deployed as a microservice-based application.

Radio Access Network

The deployment of the virtual machines implementing the CU, DU, and RIC components needs to consider their performance requirements to assure a sustained QoS on the UE connections. This implies the ability to configure and manage various low-level aspects in the virtualization layer, namely:

- Use of high performance virtual switch elements to guarantee fast processing of network packets with a minimal impact on the overall latency. Usually this functionality is provided by interconnecting the application VMs through the Open vSwitch using the Data Plane Development Kit (DPDK) datapath.

- Configure CPU and Non-Uniform Memory Access (NUMA) pinning to guarantee the resources allocated to the VM. The Edge segment consists only of a few nodes that leverage over-commitment to deploy multiple applications on a shared infrastructure. In this multi-tenant scenario it is needed to effectively isolate each workload, prioritizing those components with higher impact in the overall latency.

Additionally, it is needed to provide the virtual elements of the 5G Network with suitable configuration parameters to perform any bootstrap operation in a *zero-touch* manner.

The Mobile Core

A 5G Core performs multiple tasks like authentication and tracking of UE devices, ensuring that the IP connectivity provides the required QoS for each user and is responsible for the metering of network resources, among other things.

In OpenNebula, whenever possible we will consider mobile cores implemented as microservice applications (*cloud native)* that require a suitable platform to run a set of interconnected containers.

Complementary features implemented for the 5G-Advanced use cases

AI-based RAN load management and mobility optimization as mentioned above is a mandatory requirement that, when associated with third-party loads described in the use cases, raises the level of sophistication envisioned thus far.

C.2. 5G Edge Cluster

The deployment and operations of a highly distributed set of nodes at the edge requires specific solutions to adapt infrastructure services to the resources available at the node. This adaptation refers to both the available capabilities in terms of storage or network and the management techniques necessary to operate a large number of distributed nodes.

Orchestration and Infrastructure Management

To operate a distributed infrastructure it is usually necessary to automate the deployment and operations of each segment of the infrastructure. We'll consider the following capabilities for the orchestration layer:

- Automatic configuration of edge locations, including the installation of the required software packages and basic setup of the infrastructure including the virtual switching fabric.
- In order to properly decouple the applications with the underlying infrastructure we will use bare metal servers on the edge and cloud segments. This type of servers can run virtualized workloads without any performance penalty. When using the public cloud segment the platform should be able to use multiple providers.
- The orchestration layer should be able to determine the best location for each application component based on the application profile. Moreover, the deployment of each application component should be tailored to the distributed nature of the edge-cloud continuum.

Also, this type of infrastructure is operated by large teams that require proper mechanisms to isolate the activities which are often restricted to specific geographical locations. The orchestration layers should allow defining fine-grained access levels to each different managed object, including applications, virtual machines, and edge locations.

Edge Cluster Architecture

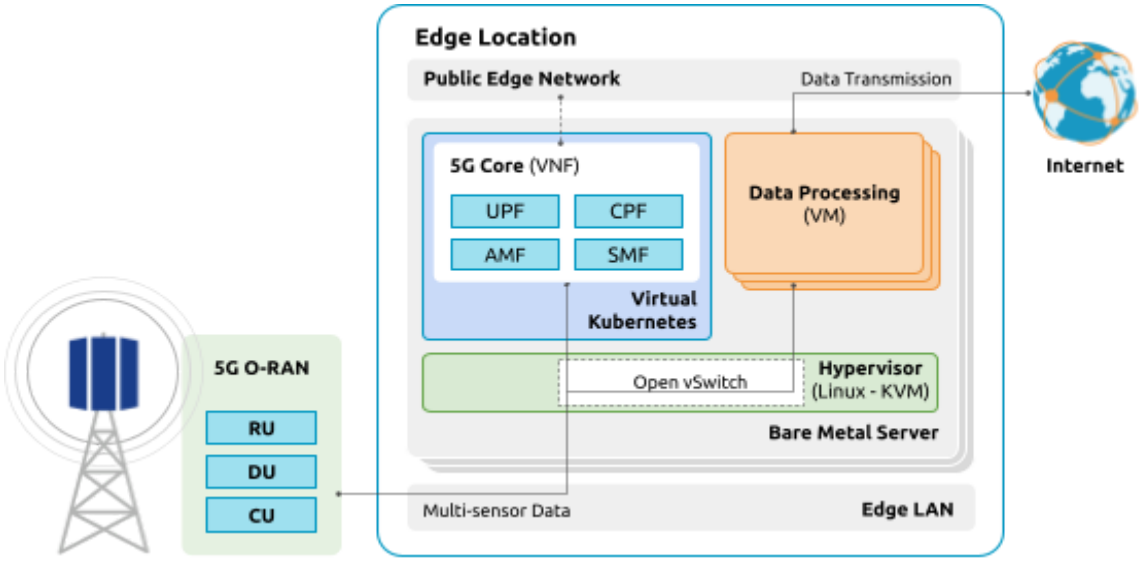


Figure 6. Main elements of the Edge Cluster.

The main element of the Edge Segment are the *Edge Clusters*. An edge cluster includes all the hardware resources needed to implement 5G networks and deploy generic applications in the cloud-edge continuum. Figure 6 presents the components of an Edge Cluster. In particular we consider the following components:

- **Storage.** It provides persistent storage to support the virtual disk images of the VMs. Considering the resource constraints of the edge locations, the storage solution needs to be lightweight while providing adequate levels of reliability and performance.

- **Networking.** The most common scenario is ToR switches in a leaf-spine configuration. The switches provide L2 connectivity for the cluster elements. It is possible to run some elements of the mobile core as P4 programs in the switches, e.g. the UPF.

- **Hypervisor.** An Edge location consists of a small number of nodes that support the execution of VMs with the KVM/QEMU hypervisor. The hypervisors also connect the VMs to the edge network through virtual switches and may expose access to additional resources like PCI or SR-IOV devices.The management, deployment, and configuration of 5G Edge Clusters must first and foremost be fully automated; additionally, it must integrate general-purpose resources (usually 2-3 servers) with telecommunications equipment (Radio Access Network, RAN). In general, most of these functions must be virtualized (Network Virtual Functions, NFV) to offer the necessary flexibility to allow the dynamic deployment of 5G applications to process traffic at the node (local break-outs).

C.3. 5G Workflows Management

The orchestration layer needs to provide a management module for the complex 5G applications consisting of multiple VM and networks. The workflow module is responsible for:

- **Management of the Application.** It includes the complete life-cycle of the application:
    - Creating all the associated components.

    - Deploying the VMs satisfying any dependencies.
    - Performing operations like terminating the application, recovering it from a failure condition, or scaling-up/down an application role.
- **Application auto-scaling.** The workflow management module evaluates the auto-scaling rules associated with each role. When a rule is violated it adjusts the cardinality of the role so the application fulfills the rule again.

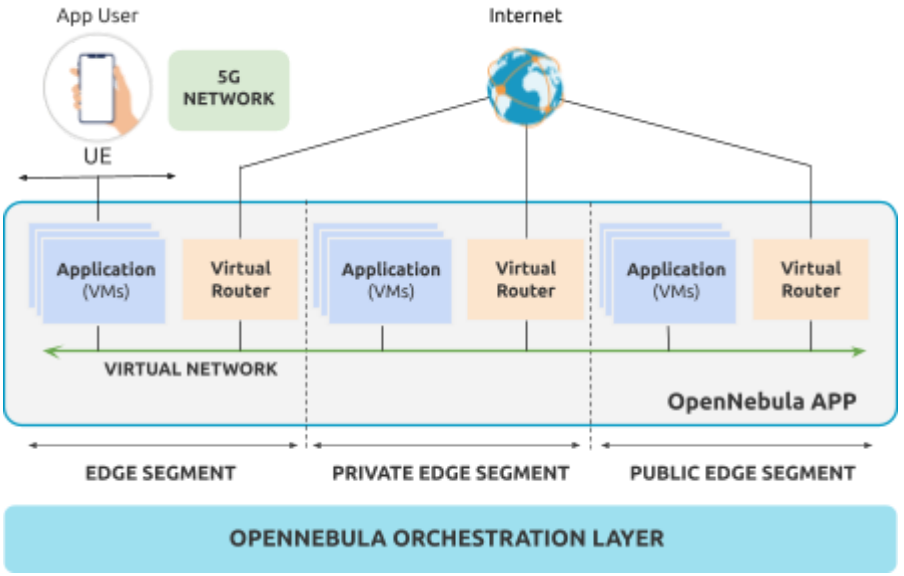


Figure 7. Architecture of an OpenNebula application running across multiple locations.

An edge-continuum application needs to be able to deploy components across multiple locations, each one located in a different segment (see description above). This requirement presents several challenges for the orchestration layer, namely:

- **Management of distributed infrastructure.** The deployment of components on hypervisors distributed across different locations requires an efficient way of distributing disk images for each application component. Additionally, the total number of resources in the edge-continuum grows linearly with the number of edge locations. The orchestration layer needs to scale its monitoring and management modules for a large number of hypervisors.

- **Virtual Networks across locations.** The virtual networks interconnecting each application component need to span multiple locations. Moreover, the application virtual networks need to be dynamically created and potentially expanded as the application needs change. Figure 7 shows the application architecture; in this case, each application segment includes a Virtual Router (VR) responsible for setting up a VPN to interconnect all application components. The VR can also provide other network functions for each segment.4. 5G Workflows AI-Driven Orchestration

The AI-Driven Orchestrator consists of four main components, as shown in Figure 8.

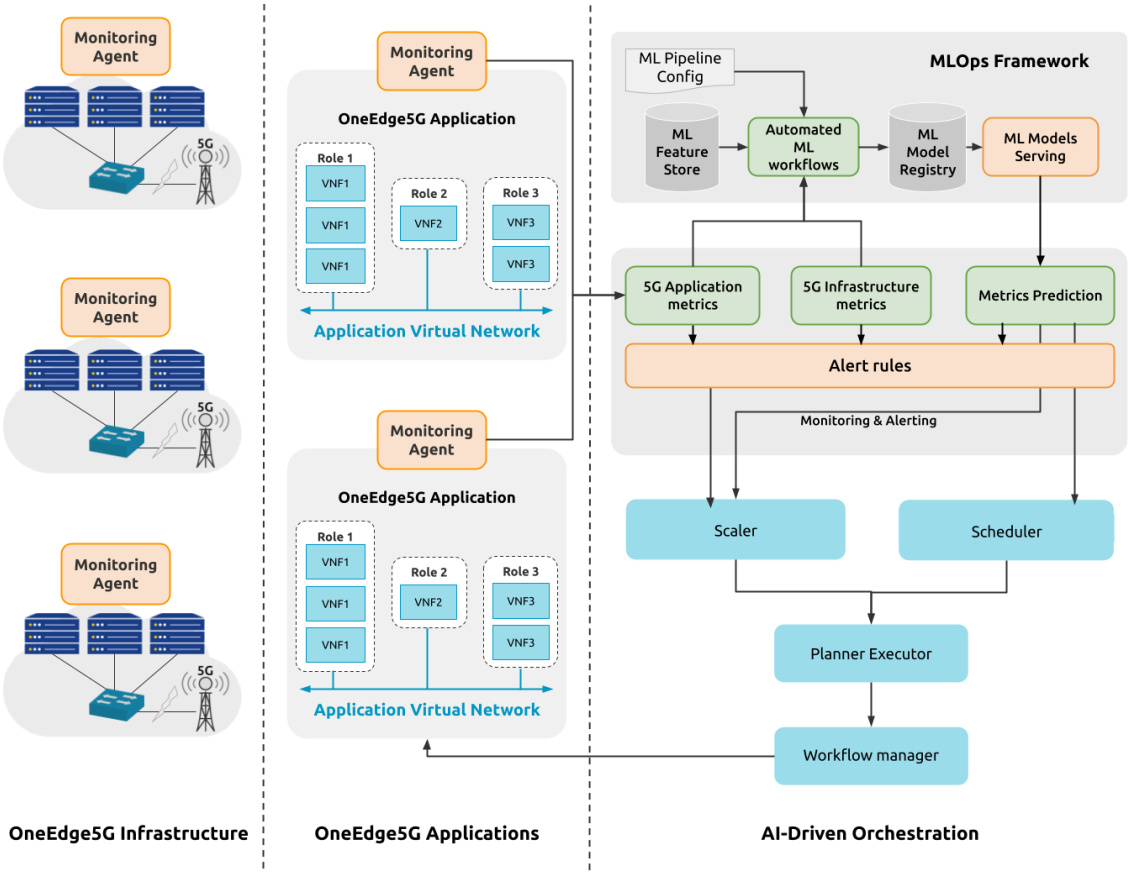


Figure 8. Interoperability of the AI-Driven Orchestrator components.

The **Monitoring & Prediction** component collects and stores metrics related to applications and Edge Cluster resources. Workflow elements (i.e. VNFs) are instrumented in order to export metrics so that the Monitoring system can actively retrieve them using a pull mechanism. Furthermore, this component provides an alerting system that allows the triggering of Orchestrator actions such as scaling based on thresholds for metrics values.

The **Scheduler** component is in charge of deploying applications on the Edge Clusters across the cloud-edge continuum, ensuring that the requirements regarding SFC latency and bandwidth are satisfied and taking into account the capacity of the physical infrastructure, administrative constraints, and optimization criteria such as energy consumption and carbon footprint.

The **Scaler** component is in charge of dynamically autoscaling 5G applications while they are running, in order to maintain application performance and service quality at predefined levels and to avoid under and over-provision. The Scaler, according to the prediction provided by the Monitoring & Prediction component, can produce plans related to:

- Vertical auto scaling, i.e. increasing/decreasing the number of CPUs, the memory, and/or the network bandwidth of a particular VNF.
- Horizontal autoscaling, i.e. increasing/decreasing the cardinality of a certain role in the OpenNebula workflow.

The auto scaling process is predictive, which means that scaling decisions are based on predictions rather than on the existing state, in order to anticipate the potential performance and quality issues and react in a timely manner. When an alert is raised by the **Monitoring and Prediction** component due to some exceeded limit threshold, the **Scaler** receives a request for scaling the corresponding

SFC workflow and it produces a plan that contains information about which roles should be scaled in or out and which VNF should be scaled up or down.

The **Plan Executor** is in charge of converting the plans produced by the **Scheduler** and the **Scaler** into a **sequence of actions** for the Workflow Manager, such as creating a VM on a new host or setting the cardinality for an application role.5. Containerized Applications Enablement

Some of the applications of interest for OpenNebula are distributed as containers to be deployed through a suitable container orchestration system, typically Kubernetes. In the project, we propose the use of a common approach that virtualizes the container application layer by running a Kubernetes cluster in a set of VMs. Each application that requires the running of containers will include a virtualized Kubernetes cluster, in particular we will use the OpenNebula Kubernetes Edition (OneKE) described below.

OpenNebula Kubernetes Edition (OneKE)

OneKE is a streamlined hyperconverged Kubernetes solution bundled with OpenNebula. Built upon RKE2 (*Rancher' s Kubernetes distribution*[2]), OneKE includes preconfigured elements for managing persistence, handling ingress traffic, and on-premises load balancing.

Figure 9 shows the architecture of a running OneKE cluster. A virtual OneKE cluster consists of the following elements:

- **Virtual Router (VR).** The virtual router connects the cluster to a common access network and provides access to the Kubernetes control plane, provides load balancing capabilities for the ingress traffic of any application running in the Kubernetes cluster.
- **Kubernetes Master Node.** This is the main orchestration component of the Kubernetes cluster and typically includes: API server that exposes the Kubernetes API to interact with the system; the Scheduler that allocates containers according to different policies and resource availability; and Controller Manager that adapts the workloads to match the desired state.
- **Kubernetes Worker Node.** The worker nodes are responsible for running the application containers (pods) and setting up the associated networking for the containers.
- **Storage Nodes.** The storage nodes provide persistent storage volumes for the Kubernetes applications. OneKE leverages the Longhorn distributed storage system to provide this functionality.

The Kubernetes cluster is also formulated as an OpenNebula application, where each component is specified as a role with a given backend VM template and cardinality, all connected through a dedicated private network. Note that the cardinality of each role can be greater than one, therefore allowing a high availability deployment for each component.

[2] https://docs.rke2.io/

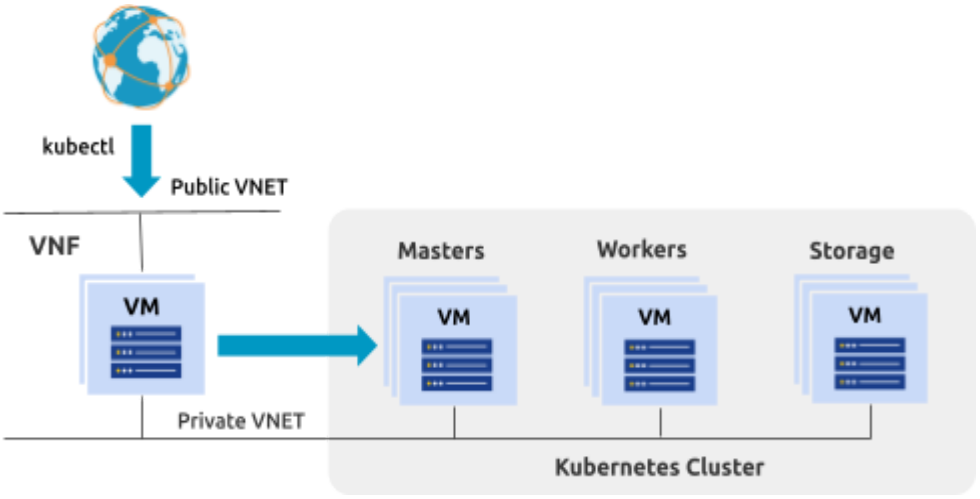


Figure 9. Architecture of a OneKE appliance.

In order to interconnect multiple Kubernetes clusters (or application components) running across multiple segments, the VR appliance needs to create a virtual private network (VPN) interconnecting all application virtual networks. The VPN mode needs to be a federated one in order to build the network overlay. Additionally, the cryptographic keys used to establish a secure connection between the routers need to be stored in a secure way in the OpenNebula framework.

Containerized Apps Integration

In order to effectively run containerized applications in a virtualized container cluster both orchestration layers (OpenNebula and Kubernetes in our case) need to be coordinated. In general this requires:

- Obtaining runtime information of the applications (pods) running in the Kubernetes cluster across all relevant namespaces.
- Obtaining the status of each worker node from both orchestration layers — Kubernetes and OpenNebula — and properly correlating them.
- The OpenNebula orchestration layer not only needs to manage the deployment of the Kubernetes virtual clusters, supervising each role and VM execution. OpenNebula also needs to push the relevant information to properly deploy the Kubernetes applications.
- The elasticity of the Kubernetes cluster needs to combine multiple information sources from the hypervisor (the host running the VM), VM (the virtual machine running the worker node), and Kubernetes (the worker node running the pods). This information has to be properly exposed to the scheduling and ML/AI modules.

Also, OpenNebula needs to expose through its API and data model a simple way to deploy applications in the Kubernetes clusters. This feature allows the deployment of a completely zero-touch enabled, hybrid application that combines virtual machines and container applications.

One of the main advantages of the application model presented in Section 2.1.1. is its ability to deploy components across the cloud-edge continuum. This also enables the seamless migration of

the containerized applications across different Kubernetes clusters in different zones. This requires integration of the multi-cluster feature of Kubernetes with the OpenNebula orchestration layer.

Finally, considering the resource constraints that some locations may have, especially 5G edge nodes, the Kubernetes cluster that runs the containers can be fitted in a single server. The use of a virtualized version (OneKE) allows us to implement this consolidation strategy.

C.6. Ready for a Test Drive?

You can evaluate OpenNebula and build a cloud in just a few minutes by using miniONE, our deployment tool for quickly installing an OpenNebula Front-end inside a Virtual Machine or physical host, which you can then use to easily add remote resources.

C.7. Conclusions

OpenNebula development takes into account the main sustainability, interoperability, and security requirements, as well as the Telco Cloud and use case requirements derived from industry-specific deployments. These requirements have shaped our product, placing it at the core of existing and new use cases, enabling AI/ML-driven orchestration for highly-distributed infrastructures and the efficient deployment and operation of the most innovative hybrid workloads.

Through the OpenNebula platform, we create a future-proof open-source solution that can be of great help in transitioning from 5G to 6G networks, addressing industrial shortcomings in the support for the cloud-edge continuum, and integrating new service capabilities — all while avoiding vendor lock-in, ensuring interoperability and service continuity, and prioritizing sustainability and efficient energy usage.

The OpenNebula model and architecture are the foundational items for innovative use-cases. Starting from its own application model, mapping onto it the 5G network model and addressing the operational challenges in application deployment and management, OpenNebula enables the development of efficient and robust digital cloud-edge continuum infrastructures able to support B5G workloads.

OpenNebula's feature set will help operators to modernize their existing networks, simplify network operations, deploy (B)5G networks quickly, integrate with upcoming 5G-Advanced and 6G features, and adopt open frameworks such as Open RAN while navigating the disaggregation of their resources

LET US HELP YOU DESIGN, BUILD, AND OPERATE YOUR CLOUD

CONSULTING & ENGINEERING

Our experts will help you design, integrate, build, and operate an OpenNebula cloud infrastructure

OPENNEBULA SUBSCRIPTION

Get access to our Enterprise Edition and to our support and exclusive services for Corporate Users

CLOUD DEPLOYMENT

Focus on your business and let us take care of setting up your OpenNebula cloud infrastructure

Acknowledgements

ONEedge5G (TSI-064200-2023-1) is supported by the Spanish Ministry for Digital Transformation and Civil Service through the UNICO I+D 6G Program, co-funded by the European Union – NextGenerationE through the Recovery and Resilience Facility (RRF)

Sign up for updates at OpenNebula.io/getupdated

Rev1.0_20250220