Telco AI Factories:
An OpenNebula Reference Architecture

Version 1.0 – February 2026

Abstract

The telecommunications industry is navigating a structural transformation of unprecedented scale, moving from a connectivity-centric model to one defined by distributed intelligence. As core connectivity revenues flatten and the commoditization of data transport accelerates, Mobile Network Operators (MNOs) are uniquely positioned to capitalize on the industrialization of Artificial Intelligence (AI). This transition requires more than merely deploying servers at the edge; it demands the creation of **AI Factories** —sovereign, highly automated, and multi-tenant infrastructures capable of executing AI models with the same reliability, deterministic performance, and security as the mission-critical network functions they host.

This white paper presents a comprehensive Reference Architecture for the OpenNebula AI Factory, specifically tailored for the telecommunications sector. It translates abstract business goals—data sovereignty, latency reduction, and infrastructure monetization—into concrete technical specifications and operational workflows. The optimal architecture for a Telco AI Factory is not a monolithic public cloud stack, but a converged, layered ecosystem that unifies High-Performance Computing (HPC) acceleration with the agility of cloud-native orchestration.

The proposed architecture addresses critical gaps in current edge deployments, specifically the lack of seamless integration between 5G User Plane Function (UPF) and AI inference engines, the challenge of telemetry normalization for AI models, and the operational complexity of managing heterogeneous workloads (vRAN + AI) on shared silicon. By leveraging OpenNebula's zero-overhead virtualization, integrated with advanced networking fabrics and automated service orchestration, operators can transform their distributed footprint into a sovereign AI-ready engine that serves the diverse needs of the European industrial landscape.

Contents

Acronyms and Abbreviations

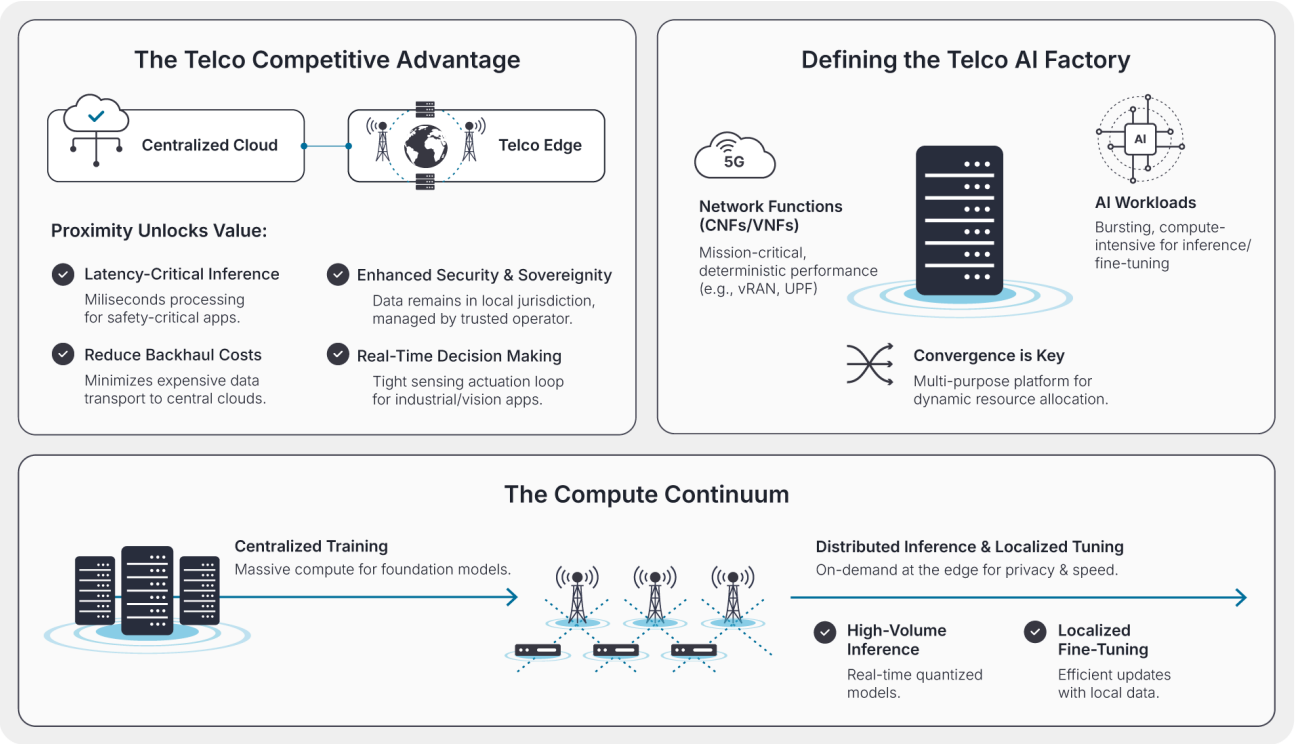
ACL	Access Control List
AI	Artificial Intelligence
CNF	Cloud-Native Network Function
CU	Centralized Unit
DNN	Data Network Name
DPU	Data Processing Unit
DU	Distributed Unit
EPA	Enhanced Platform Awareness
GPU	Graphics Processing Unit
HPC	High-Performance Computing
IaaS	Infrastructure-as-a-Service
LLM	Large Language Model
LRMs	Large Reasoning Models
MIG	Multi-Instance GPU
MLOps	Machine Learning Operations
MNO	Mobile Network Operator
NIC	Network Interface Card
NUMA	Non-Uniform Memory Access
SLMs	Small Language Models
SR-IOV	Single Root I/O Virtualization
UMA	Unified Memory Architecture
UPF	User Plane Function
VDC	Virtual Data Center
VM	Virtual Machine
VNF	Virtual Network Function
vRAN	Virtualized Radio Access Network

1. The Rise of the Telco AI Factory

Telecommunications operators are currently facing a pivotal market shift where traditional connectivity revenue is maturing, while demand for localized, secure computing is growing at approximately 20-25% annually. Market research by STL Partners, for example, indicates that by 2028 telcos will have deployed more than 1800 network edge sites, representing a massive expansion of distributed compute capacity.[1] While early edge deployments focused primarily on reducing latency for consumer applications like gaming, today the primary driver for modern enterprise Artificial Intelligence (AI) has shifted toward **data residency, sovereignty, and operational resilience**.

The industrial landscape is increasingly dominated by applications that require real-time processing and strict data governance, such as industrial automation, autonomous mobility, and real-time video analytics for public safety and industry. These applications generate petabytes of data that are often too voluminous, sensitive, or latency-critical to transport to centralized hyperscale clouds for processing.

[1] https://stlpartners.com/press/2000-network-edge-data-centres-by-2028/



The Telco Competitive Advantage

While hyperscale cloud providers dominate centralized model training, Mobile Network Operators (MNOs) possess a distinct competitive advantage: a dense, capillary infrastructure capable of running compute-intensive AI workloads closer to the source of data than any centralized provider. This proximity unlocks four critical value propositions:

1. **Latency-Critical Inference:** Processing data within milliseconds of its generation, essential for safety-critical applications.
2. **Reduced Backhaul Costs:** Processing data at the edge minimizes the expensive transport of large datasets to central clouds.
3. **Enhanced Security and Sovereignty:** Ensuring sensitive data remains within a specific jurisdiction or even on-premise, managed by a trusted local operator, which is a key requirement for government and regulated industries.
4. **Real-time Decision Making:** Co-locating AI inference with network functions enables low-latency responses for industrial and vision-based applications, creating a tight feedback loop between sensing and actuation.

Defining the Telco AI Factory

A Telco AI Factory differs fundamentally from a standard cloud data center. It is not a general-purpose hosting environment but a specialized facility designed to manufacture intelligence. It must host two distinct types of workloads on the same physical infrastructure:

1. **Network Functions (CNFs/VNFs):** Mission-critical workloads like the 5G Virtualized Radio Access Network (vRAN) and 5G Core User Plane Function (UPF). These require deterministic performance, precise timing, and direct hardware access to maintain network stability.
2. **AI Workloads:** Bursting, resource-intensive applications for inference and fine-tuning, like Large Language Models (LLMs), Small Language Models (SLMs), Large Reasoning Models (LRMs), and Computer Vision, that require dynamic access to distributed, accelerated hardware capabilities.

OpenNebula addresses this convergence by providing a unified control plane that orchestrates both workload types while ensuring that noisy AI applications do not degrade the mobile network's performance. For Telcos, this offers a dual value proposition:

- **Internal Optimization (AI for Telco):** Using AI to optimize spectral efficiency, beamforming, and power consumption within the radio network itself, reducing OPEX.
- **External Monetization (AI on Telco):** Selling excess Graphics Processing Unit (GPU) capacity at the edge to enterprise customers for workloads requiring low latency or strict data residency, generating new revenue streams.

Historically, Telcos maintained separate silos for network functions (RAN/Core) and IT workloads. The AI Factory requires a new convergence. To be economically viable, the edge node must be a multi-purpose environment. It must run the 5G vRAN software during the day to support mobile subscribers and AI for Telco demands via dynamic partitioning mechanisms, and during off-peak hours utilize the same GPU resources to run AI training or inference tasks.

The Compute Continuum

A critical architectural consideration for the Telco AI Factory is the balance between centralized training capabilities and distributed inference requirements. While training will be typically carried out in the cloud or on supercomputers due to the massive computational and energy requirements of large-scale foundation models, **AI inference will primarily occur on-demand at the telco edge**.

Moreover, we must consider that, in some cases, it may be necessary to perform not only inference but also **distributed training of foundation models at the edge** to meet data privacy and security requirements. In highly regulated sectors or for defense applications, for example, exporting sensitive data to a central cloud for fine-tuning might not be admissible.

In these cases, resources are constrained, so a balance must be struck between the model size (to fit edge resources) and its accuracy. The edge infrastructure must therefore be versatile enough to support:

- **High-Volume Inference:** Running quantized models for real-time decision-making.
- **Localized Fine-Tuning:** Utilizing efficient techniques (e.g. Low-Rank Adaptation) to update model weights based on local data without requiring the massive infrastructure of pre-training.

2. Technical Requirements for Telco AI Inference

To operate as a carrier-grade AI platform, the architecture must satisfy specific technical requirements for telecommunication environments. Unlike standard cloud environments where resources are often overcommitted to maximize density, telco AI workloads require precise control over hardware allocation to avoid performance degradation and ensure the deterministic behavior required by 5G standards.

Requirement 1: Deterministic Performance via Enhanced Platform Awareness

The coexistence of real-time network functions (like vRAN) and bursty AI workloads demands an infrastructure capable of eliminating noisy neighbor interference. The virtualization layer must support **Enhanced Platform Awareness (EPA)** to provide granular control over how virtual resources map to physical hardware.

- **Deterministic CPU Allocation:** The platform must support strict **CPU Pinning** policies. This requirement prevents the host operating system's scheduler from migrating virtual CPUs (vCPUs) between physical cores, which would otherwise cause context-switching overhead and cache invalidation. For latency-sensitive inference, a vCPU must be essentially a dedicated physical core.

- **Memory Topology Optimization:** Modern servers use Non-Uniform Memory Access (NUMA) architectures. The orchestration layer must be **NUMA-aware**, ensuring that a workload's compute processes and memory allocations are localized to the same CPU socket. Accessing memory across the inter-socket interconnect incurs unacceptable latency penalties for 5G signal processing or real-time inference.
- **Memory Management Efficiency:** To support LLMs with massive memory footprints, the platform must support **HugePages**. This reduces memory management overhead, ensuring efficient memory addressing for large datasets.
- **Hardware Passthrough:** For maximum performance, the architecture must support Single Root I/O Virtualization (**SR-IOV**) and **PCI Passthrough**. This requirement allows Virtual Machines (VMs) and containers to bypass the hypervisor kernel entirely and communicate directly with GPUs and Network Interface Cards (NICs), eliminating the virtualization I/O tax.

Requirement 2: Strict Multi-Tenancy and Secure Isolation

A Telco AI Factory is inherently a multi-tenant environment, hosting internal network functions alongside external enterprise workloads. The architecture must enforce a hard security boundary between these tenants to prevent data leakage and resource contention.

- **Logical and Physical Segregation:** The system must support the partitioning of resources into **Virtual Data Centers (VDCs)**. A VDC acts as a self-contained logical cloud where a tenant has exclusive access to a defined quota of compute, storage, and networking resources. Crucially, the resource consumption of one VDC must not impact the performance of another.
- **Granular Governance:** Governance must be enforced via fine-grained **Access Control Lists (ACLs)** that manage permissions for different user personas (e.g., infrastructure admins, service admins, end users). This governance model ensures that a tenant can manage their own services without visibility into the underlying physical infrastructure or other tenants' environments.
- **Secure Networking Boundaries:** Security enforcement (firewalling, encryption) should be offloaded to specialized hardware, such as **Data Processing Units (DPUs)**. This creates a physical separation between the tenant workloads running on the host CPU and the security policies enforced on the network card, significantly hardening the attack surface.

Requirement 3: Heterogeneous Execution Models

The diversity of AI workloads means a homogeneous execution model is insufficient. The architecture must support multiple modes of operation to cater to different stages of the AI lifecycle and different user technical maturities.

- **Cloud-Native Orchestration (Kubernetes-based):** This model is suited for multi-user setups, MLOps pipelines, and scalable microservices. Here, the platform must orchestrate the lifecycle of Kubernetes clusters, which in turn manage containerized AI applications.
- **High-Performance Direct Execution (HPC-styled):** For performance-critical inference and heavy training jobs, the overhead of container networking and orchestration layers can be prohibitive. The architecture must support a "bare-metal-like" execution mode where workloads run directly on optimized VMs with GPU passthrough. This approach provides consistent latency and predictable behavior, essential for foundation model fine-tuning and AI-RAN signal processing.

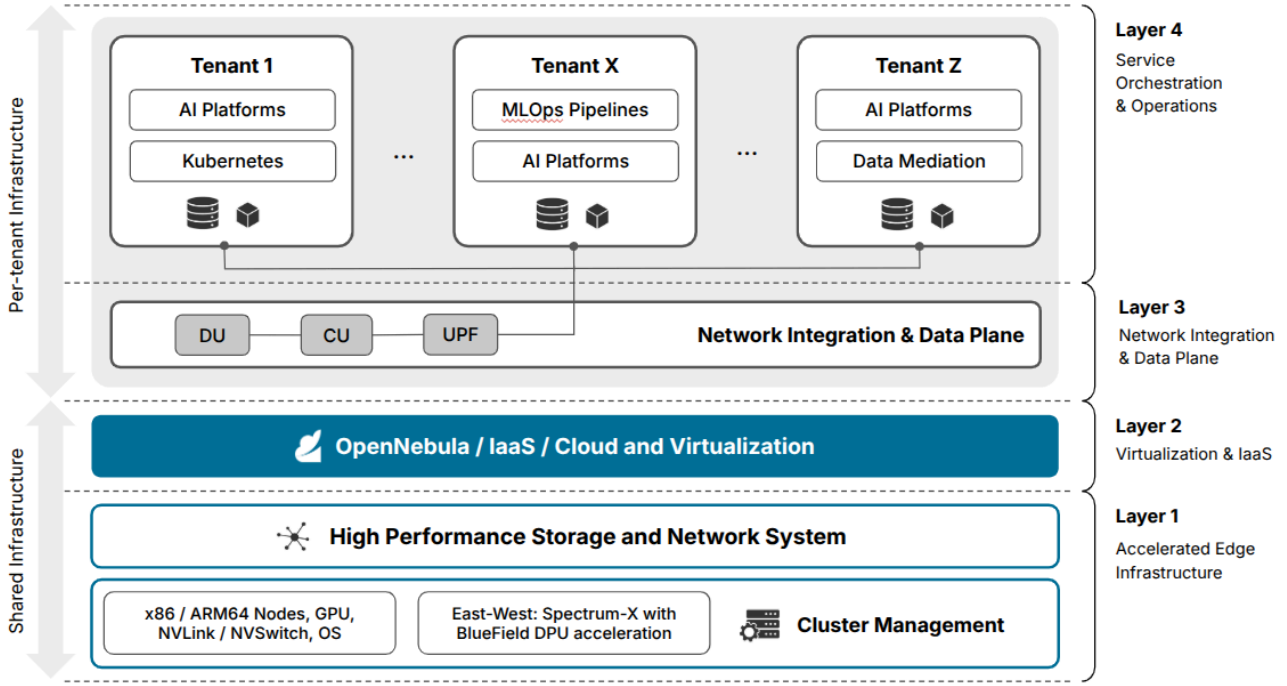
3. Reference Architecture Overview

The **OpenNebula AI Factory Reference Architecture**[2] is structured in layers, establishing a clear functional separation between the management of shared physical infrastructure and the management of different

[2] https://support.opennebula.pro/hc/en-us/articles/31630339592861-White-Paper-OpenNebula-AI-Factory-Reference-Architecture

service delivery platforms per tenant. This model ensures full isolation, scalability, security, and the ability to support a multi-vendor ecosystem, while allowing telcos to scale up their AI offerings without increasing operational complexity.

- **Layer 1: Accelerated Edge Infrastructure.** The physical foundation consists of distributed edge nodes, GPU-accelerated servers, and high-performance interconnection devices (SmartNICs/DPUs) and storage systems.
- **Layer 2: Virtualization & Infrastructure-as-a-Service (IaaS).** The OpenNebula orchestration plane that abstracts physical resources, manages the lifecycle of VMs and Kubernetes clusters, and enforces strict isolation.
- **Layer 3: Network Integration & Data Plane.** The critical integration layer handles data-plane traffic via 5G UPF traffic steering and Local Breakout.
- **Layer 4: Service Orchestration & Operations.** The automation layer is responsible for deploying complex, multi-component services and managing the data mediation and AI lifecycle (MLOps).



4. Accelerated Edge Infrastructure Layer

The foundation of the AI Factory is specialized hardware. At the base of the AI Factory, the **Accelerated Edge Infrastructure Layer** provides high-performance hardware tailored for telco environments. This physical foundation consists of distributed edge nodes equipped with hardware acceleration and specialized networking fabrics and storage systems designed to handle the massive throughput required by AI workloads while adhering to the strict power, space, and environmental constraints of the network edge.

Heterogeneous Compute Architectures

To address the diverse requirements of training (throughput) and inference (latency/efficiency), the architecture supports a **heterogeneous mix of node types**.

- **High-Performance x86 Nodes:** Standard enterprise servers equipped with high-end PCIe Gen5 buses to support full-sized GPUs. These nodes provide the optimized performance necessary for complex model fine-tuning, unifying memory, and massive parallel processing.
- **Energy-Efficient ARM64 Nodes:** For far-edge deployments such as street cabinets or cell towers where power and cooling are constrained, the architecture leverages ARM-based processors.

OpenNebula's native support for ARM ensures a unified management plane across x86 and ARM resources, allowing operators to deploy workloads to the most efficient architecture seamlessly.

- **Unified Memory Architectures (UMA):** This architecture allows the use of a unified CPU and GPU memory space via high-speed interconnects, eliminating the PCIe bottleneck. This is crucial for running LLMs at the edge, as it allows models that exceed standard GPU VRAM capacities to access system memory with high bandwidth, enabling the deployment of larger, more accurate models without requiring a massive data center footprint.

The High-Performance Interconnect Fabric

In an AI Factory, the network is effectively the computer. East-west traffic—communication between nodes during distributed training—requires bandwidth and latency characteristics that traditional TCP/IP cannot provide.

- **Spectrum-X Ethernet:** We recommend NVIDIA Spectrum-X as the standard fabric. It enables RDMA over Converged Ethernet extensions optimized for AI workloads, providing the low latency and lossless behavior required for AI, while retaining the manageability of standard Ethernet. This ensures that GPU clusters can communicate without network-induced bottlenecks, which is vital for multi-node training jobs.
- **Data Processing Units:** The inclusion of DPUs (e.g., NVIDIA BlueField-3) is a mandatory requirement for the secure AI Factory. The DPU offloads the virtualization layer (Open vSwitch, security groups, firewalling) from the host CPU. This serves two critical functions: it frees up 100% of the host CPU for revenue-generating workloads (vRAN or AI), and it creates a hard security boundary by running the infrastructure control plane on separate silicon from the tenant workloads. This isolation is a prerequisite for hosting untrusted third-party code alongside critical network functions.

Distributed Storage Systems

Storage at the edge must solve a paradox: it requires the high throughput of a parallel file system for training (sequential reads) and the low latency of block storage for inference (random reads).

- **Local NVMe Tier:** Local high-speed storage used for ephemeral caching and scratch space during inference tasks, ensuring minimal latency for model loading and context retrieval.
- **Distributed Storage Fabric:** The architecture leverages software-defined storage solutions or specialized high-performance file systems (e.g., WEKA, VAST Data, NetApp) integrated via native drivers. These systems provide a unified namespace across the edge cluster, allowing data to be accessed concurrently by multiple GPU nodes at high throughput.
- **Global Data Lake Integration:** This edge storage could function as a caching tier, synchronizing efficiently with a cloud data lake. This would allow the AI Factory to pull model weights or reference data from the cloud only when necessary, minimizing backhaul bandwidth consumption. While local storage ensures data sovereignty, this optional integration would enable hybrid workflows where models are trained centrally and deployed to the edge.

5. Virtualization & Infrastructure-as-a-Service Layer

This layer provides the Operating System for the AI Factory. At the **Virtualization & IaaS Layer**, OpenNebula acts as the centralized manager providing unified orchestration across distributed edge clusters with zero-overhead virtualization.

Zero-Overhead Virtualization Strategy

A Telco AI Factory cannot afford the 5-10% performance penalty typical of generic hypervisors.

OpenNebula eliminates this by configuring KVM technologies for near-native performance.

- **PCI Passthrough & SR-IOV:** The architecture mandates the use of SR-IOV for NICs and full PCI Passthrough for GPUs. This allows the VM or container to bypass the hypervisor kernel entirely and communicate directly with the hardware. This is critical for AI-RAN workloads where microseconds matter.
- **CPU Pinning & NUMA Awareness:** To safely co-locate vRAN and AI, the orchestration layer must map specific virtual vCPUs to specific physical cores. OpenNebula's scheduler is configured to be NUMA-aware, ensuring that memory allocated to a VM resides on the same memory node as the CPU cores it is using. This prevents the latency jitter caused by crossing the interconnect between sockets, which is fatal for real-time RAN processes.
- **Negative Overhead:** Contrary to the belief that virtualization degrades performance, research[3] indicates that properly configured virtualized AI infrastructure can yield negative overhead compared to an untuned bare-metal environment. By using HugePages, CPU Pinning, and interrupt affinity tuning within the KVM hypervisor templates, AI workloads can run more consistently than under a general-purpose OS scheduler. This deterministic behavior is crucial for 5G vRAN and real-time inference, effectively providing a cleaner, more stable execution environment.

Multi-Tenant Governance Model

Secure multi-tenancy is intrinsic to the architecture, not just a feature. The architecture uses OpenNebula's native Virtual Data Center (VDC) construct to partition the physical cluster.

- **Resource Isolation:** A tenant (e.g., a particular private customer) is assigned a VDC, which acts as a logical cloud. They see only their quota of compute, storage, and networking resources. This abstraction allows the physical infrastructure to be shared while ensuring resource allocation guarantees.
- **Access Control Lists:** Fine-grained permissions determine who can deploy services, view monitoring data, and manage the underlying infrastructure. This allows a hierarchical model where the Telco manages the physical layer, a Managed Service Provider manages the VDC, and the end-user deploys the application.
- **Marketplace Integration:** Tenants have access to a curated set of appliances providing pre-configured images of AI frameworks (e.g., vLLM). This simplifies the user experience, allowing them to deploy a full AI stack with a single click, abstracting the complexity of the underlying drivers and libraries.
- **Catalogue of AI Models:** Tenants have access to public AI models hosted in online collections (e.g., via HuggingFace and NVIDIA NIM) but are also able to host their own catalogue with a set of private AI models that have been trained or fine-tuned in their individual AI environment.

Convergence of Execution Models

The architecture supports a dual-mode execution strategy, as not all AI workloads fit into containers.

- **Cloud-Native AI (Kubernetes-on-IaaS):** OpenNebula provisions and manages the lifecycle of Kubernetes clusters inside VMs. This provides the standard API developers expect for orchestration while maintaining hard isolation between tenants via the VM layer. Suitable for collaborative development pipelines, MLOps, and scalable microservices.
- **HPC-styled AI (Bare-Metal/VM):** Tenants bypass the Kubernetes layer and run directly on optimized VMs with passed-through GPUs. This reduces the number of abstraction layers to the absolute minimum. Suitable for massive model training, ultra-low latency inference, and AI-RAN signal processing, where the overhead of container networking is unacceptable.

[3] https://research.ibm.com/publications/to-virtualize-or-not-to-virtualize-ai-infrastructure-a-perspective

6. Network Integration & Data Plane Layer

For an AI Factory to process telco network data (e.g., user traffic from mobile devices), there must be a defined interface between the 5G Core and the AI workloads. This layer defines the **Local Breakout topology**, enabling the AI Factory to act as a data processor for the network.

The seamless integration of the 5G User Plane and the AI Compute Plane is achieved via the N6 Interface. This is the demarcation point where IP packets leave the telco tunneling protocol (GTP-U) and enter the Data Network (DN).

- **Ingress (N3 Interface):** User traffic originates from the User Equipment, traverses the RAN, and arrives at the Edge UPF via the N3 interface. The UPF is a VNF or CNF running on the same OpenNebula cluster or an adjacent node.
- **Traffic Steering & Breakout:** The Edge UPF inspects the packet headers. Based on traffic steering rules (defined by the SMF/PCF), traffic destined for the internet is routed to the backhaul. However, traffic matching a specific AI Application (identified by IP range, Data Network Name (DNN), or Application ID) is routed locally via the N6 Interface.
- **The N6 Handoff:** The N6 interface is not just a cable; in a virtualized environment, it is a logical connection. The architecture maps the N6 output of the UPF container/VM to a specific OpenNebula Virtual Network (VNet). This VNet is an isolated VLAN or overlay network dedicated to that specific tenant.
- **Ingestion & Processing:** The AI Inference VM (or Kubernetes Ingress) is attached to this specific VNet. It receives the raw IP packets directly from the UPF with minimal latency (sub-millisecond). By using SR-IOV for this handoff, the packets move directly from the NIC to the AI application's memory space, avoiding the host CPU entirely and enabling real-time processing loops crucial for applications like autonomous driving or industrial robotics.

7. Service Orchestration & Operations Layer

Lastly, at the **Service Orchestration & Operations Layer**, operators can securely host third-party AI applications, thereby monetizing the infrastructure. The complexity of the Edge AI Factory lies not just in the hardware, but in the operational management of thousands of distributed nodes. The architecture relies on OpenNebula's orchestration capabilities to manage the complex interplay between services.

Service Composition and VDC Self-Service

An AI Service is rarely a single VM. It is a composite service consisting of multiple roles (e.g., Ingress Controller, Inference Engine, Data Store). The orchestration engine manages this dependency graph, handling startup order and elasticity.

- **Service Templates:** The architecture defines services using JSON/YAML templates. These templates specify the roles, the number of VMs per role, and the startup order (e.g., start the database before the inference engine).
- **Elasticity & Auto-Scaling:** The orchestration engine monitors infrastructure metrics (e.g., GPU utilization) and application metrics (e.g., N6 throughput). It defines elasticity policies to automatically instantiate new GPU VMs during traffic bursts, ensuring that SLAs are met without manual intervention.
- **Self-Service Portals:** B2B customers can provision their own Kubernetes clusters or AI environments within their VDC without visibility into the underlying telco network. This Self-Service capability is essential for scaling the "AI on Telco" business model, allowing the operator to act as a platform provider rather than a managed service provider.

Data Engineering Foundation and Mediation

This layer also facilitates the introduction of data mediation services aligned with the **European Telco AI Platform** envisioned by the European Commission's Apply AI Strategy,[4] which calls for the integration of open source AI stack elements, including mediation layers and data engineering pipelines.

- **Data Mediation Services:** Telco networks generate massive volumes of telemetry data that are unintelligible to standard AI models. This calls for Data Mediation Services—lightweight containers that ingest raw telemetry, sanitize it, and normalize it into AI-friendly formats.
- **Telemetry Ingestion & Normalization:** Raw telemetry from routers and base stations is often vendor-specific and verbose. The mediation service normalizes this heterogeneous data into a common schema suitable for AI consumption. This data plumbing is critical for training reliable network automation models, ensuring that the AI sees a consistent view of the network regardless of the underlying vendor hardware.

Day 2 MLOps Pipeline

The AI Factory must support the continuous lifecycle of AI models (MLOps). This layer conceives the integration with MLOps platforms to form a closed loop.

- **Training (Core Cloud):** New models are trained in a centralized location or a specialized Training VDC using aggregated data.
- **Registry (Distribution):** Trained models are pushed to a container registry or Model Registry.
- **Distribution (Edge Push):** When a service update is triggered, the new model weights are pre-cached to the edge node storage systems during off-peak hours to avoid network congestion.
- **Cutover (Rolling Update):** The platform performs a rolling update, sequentially replacing old inference nodes with new ones, ensuring zero downtime for the end service.

8. Use Cases and Customer Workflows

The flexibility of the OpenNebula Telco AI Factory supports diverse workflows that drive revenue and operational efficiency. The following workflows illustrate how the architecture components interact to deliver value in real-world scenarios.

Use Case A) Sovereign AI Training and Inference Loop

In this workflow, a regulated entity (e.g., a National Health Service or Defense Agency) requires an isolated environment to fine-tune an LLM using sensitive local data.

- **Tenant Provisioning.** The customer requests a Secure AI VDC with specific GPU quotas (e.g., 8x NVIDIA H100s) and storage requirements.
- **Environment Instantiation.** OpenNebula automatically provides the requested resources and establishes a private VDC. Crucially, it configures the network to be air-gapped from the public internet, accessible only via a secure private 5G slice.
- **Data Ingestion.** Sensitive data is ingested directly from the customer's on-premise equipment via the Telco's private network (Local Breakout), bypassing public internet gateways entirely.
- **Fine-Tuning.** The customer deploys their training job using a pre-configured AI appliance from the OpenNebula Marketplace. The job runs on bare-metal capability via PCI Passthrough, ensuring maximum performance.
- **Inference Deployment.** Once the model is fine-tuned, its weights are saved to local storage. Inference endpoints are then deployed within the same VDC for local, low-latency access by the customer's applications, creating a closed-loop sovereign AI ecosystem.

[4] https://digital-strategy.ec.europa.eu/en/policies/apply-ai

Use Case B) AI-RAN Convergence

This internal workflow illustrates the operator using the AI Factory to enhance its own network performance, a concept known as AI for RAN.

- **Signal Processing.** The Radio Unit receives RF signals and forwards them to the Distributed Unit (DU) hosted on the edge node.
- **AI Acceleration.** Instead of traditional algorithms, the DU uses a neural network (Neural Receiver) to process the signal. This workload is offloaded to the same GPU that is hosting third-party AI services.
- **Resource Partitioning.** The architecture utilizes Multi-Instance GPU (MIG) technology to strictly partition the GPU. The Network Slice of the GPU guarantees compute cycles for the vRAN while utilizing the remaining capacity for best-effort B2B inference tasks.
- **Continuous Learning.** Performance data from the RAN is fed back into a training loop, optimizing the Neural Receiver model, which is then redeployed to the edge without interrupting service.

Use Case C) Real-Time Industrial Analytics

A manufacturing company deploys 5G-connected cameras in their smart factories for automated defect detection along its high-speed production lines.

- **Traffic Capture.** Cameras stream high-definition video over the private 5G network.
- **Local Breakout.** The 5G UPF at the edge site identifies the video stream and routes it locally to the specific tenant within the AI Factory, avoiding the backhaul network.
- **Ingestion and Pre-processing.** A mediation container receives the stream, decodes the video, and frames it for analysis.
- **Inference.** A computer vision model running in a Kubernetes pod (managed by OpenNebula) detects defects in real-time (<10ms).
- **Actuation.** The system sends a low-latency stop command back to the manufacturing robot, keeping the entire data loop within the AI Factory. This ultra-low-latency response is only possible because the UPF and the AI inference engine are co-located.

9. Conclusions and Next Steps

The transition to an AI-native architecture is a strategic imperative for Telecommunications operators. The OpenNebula AI Factory Reference Architecture for Telcos offers a proven, vendor-neutral path to achieving this transformation. By moving beyond simple connectivity and embracing a converged, sovereign, and intelligent infrastructure, MNOs can secure a pivotal role in the AI value chain.

This architecture allows Telcos to handle demanding, sensitive, and latency-critical AI workloads. By integrating the high-performance capabilities of accelerated hardware (Layer 1) with the zero-overhead virtualization of OpenNebula (Layer 2), the seamless connectivity of data plane traffic (Layer 3), and the automated service orchestration (Layer 4), operators can build a factory that not only processes data but generates sustainable value. This solution empowers operators to reclaim control over their distributed data assets, optimize their own networks through AI-RAN, and provide a robust, sovereign platform for the industrial AI revolution.

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, 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

Sign up for updates at OpenNebula.io/getupdated

Rev1.0_20260225