The AI Factory Reference Architecture: Building Secure, Scalable, and Sovereign AI Infrastructure with OpenNebula

Version 1.0 – November 2025

Abstract

The growing demand for artificial intelligence is driving a fundamental shift in how computing infrastructure is designed and delivered. Around the world, **cloud providers, HPC centers, telecommunications companies, and public institutions** are investing in the creation of **AI Factories** — large-scale infrastructures built to train, deploy, and operate AI models efficiently and securely.

This white paper introduces the **OpenNebula AI Factory Reference Architecture**, a comprehensive framework for building **secure, multi-tenant AI infrastructure** that unifies high-performance computing, virtualization, and cloud orchestration within a single platform. Designed to meet the growing demands of **AI training, inference, and service delivery**, this architecture establishes a clear separation between **shared infrastructure management** and **per-tenant operations**, enabling organizations to deliver scalable **AI-as-a-Service** models with full data sovereignty and near-native performance.

At its foundation, the AI Factory integrates **GPU-accelerated servers** (e.g., NVIDIA H100, GB200, GH200) interconnected via **NVLink**, **InfiniBand**, or **Spectrum-X Ethernet**, complemented by **BlueField-3 DPUs** for secure networking and **high-performance distributed storage systems** for low-latency data access. On top of this hardware layer, **OpenNebula** acts as the **Infrastructure-as-a-Service (IaaS)** layer, providing unified orchestration, virtualization, and automation capabilities across distributed clusters.

Each tenant operates within an **isolated Virtual Data Center (VDC)**, managing its own AI platforms and frameworks—ranging from Kubernetes-based environments (e.g., Run:ai, Kubeflow, NVIDIA Dynamo) to bare-metal GPU execution for performance-critical workloads. OpenNebula's **multi-tenant model** ensures complete resource isolation through **ACL-based governance**, while supporting chargeback, monitoring, and marketplace integration for self-service operations and flexible business models.

The white paper also details the **key building blocks** of the AI Factory ecosystem—including AI platforms, storage, monitoring, security, DevOps, and cluster management tools—alongside a **four-layer user model** that distinguishes between end users, service users, service administrators, and infrastructure administrators.

Finally, it demonstrates how OpenNebula achieves **near-zero virtualization overhead** through advanced technologies such as **PCI passthrough**, **SR-IOV**, and **CPU pinning**, ensuring that AI workloads and HPC jobs run with **bare-metal performance** while maintaining the agility, scalability, and isolation required for modern, multi-tenant AI infrastructures.

Contents

Acronyms and Abbreviations

AD	Active Directory
CMDB	Configuration Management Database
CPU	Central Processing Unit
DC	Datacenter
DPDK	Data Plane Development Kit
GPU	Graphics Processing Unit
HA	High Availability
HCI	Hyper Converged Infrastructure
LUN	Logical Unit Number
NAS	Network Attached Storage
NFS	Network File System
NIC	Network Interface Card
NUMA	Non-Uniform Memory Access
RADOS	Reliable Autonomic Distributed Object Storage
SAN	Storage Area Network
VDC	Virtual Datacenter
VM	Virtual Machine
VNC	Virtual Network Computing
VNF	Virtual Network Functions

1. AI Factory Reference Architecture

The AI Factory reference architecture establishes a clear separation between the management of the shared infrastructure and the management performed by each tenant over its assigned, isolated partition.

From the bottom up, **the shared infrastructure** includes:

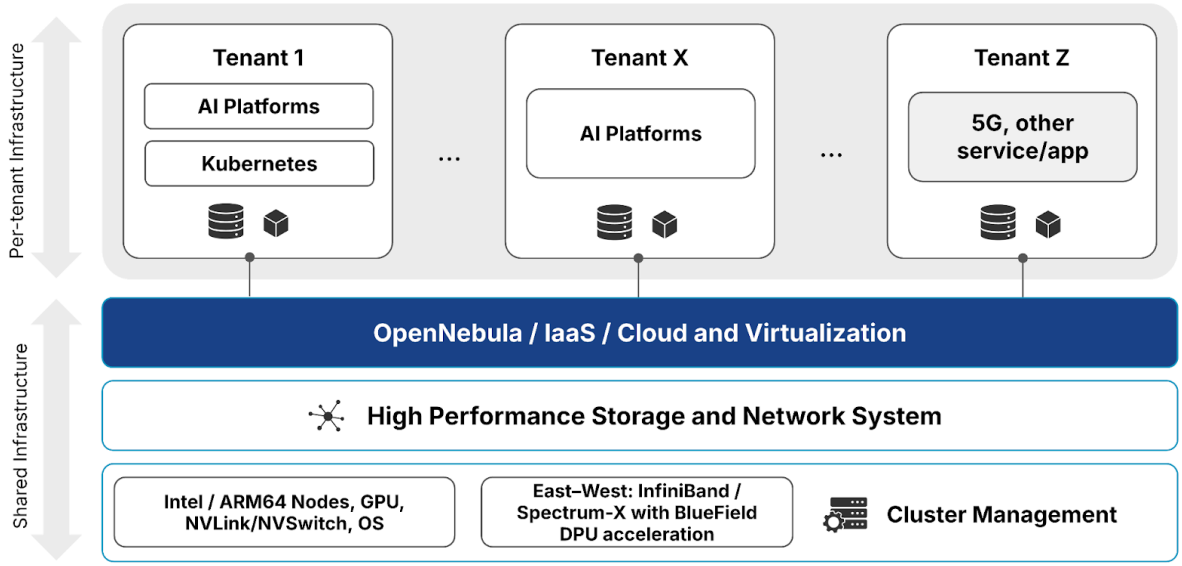
- **High-performance accelerated hardware,** based on GPU servers (e.g., NVIDIA H100, GB200, or GB300), interconnected via NVLink for multi-GPU execution. East–west data traffic is handled through InfiniBand or Ethernet Spectrum-X fabrics, enhanced with NVIDIA BlueField-3 (BF3) DPUs for secure, low-latency data processing, network offloading, and isolation between tenants and workloads.

- A **high-performance distributed storage system**, essential to enable efficient data sharing and low-latency access to datasets across AI workloads and tenants. On top of this hardware foundation,

- The **Infrastructure-as-a-Service (IaaS) layer**, providing unified orchestration, virtualization, and resource partitioning capabilities that support multi-tenancy, scalability, and automation across the AI Factory environment. All virtualized services share a common high-performance network and storage subsystem, directly connected to virtual machines through passthrough technologies to ensure low latency and high throughput.

The **per-Tenant infrastructure** includes:

- The **Virtual Data Center** (VDC) that contains the resources allocated to the Tenant for running their AI platform instances.
- Within these environments, teams and departments operate **AI development and execution platforms** to design, manage, and monitor AI workloads, as well as to deploy model endpoints for end users.
- Most AI platforms require a **Kubernetes-based environment** to manage containerized workloads efficiently. Each Tenant dynamically assigns GPU resources to its frameworks, which then distribute them across workloads according to job scheduling policies, ensuring optimal GPU utilization, workload prioritization, and overall operational efficiency.
- However there are Tenants that prefer to use frameworks and runtimes for executing AI workloads **directly on GPU servers** without an intermediate Kubernetes layer, offering maximum performance for specific inference or training use cases.



2. HPC AI and Cloud-Native AI

Kubernetes has become a powerful tool for managing containerized workloads, but on its own, it's not enough to build a full AI Factory. Running large-scale AI systems requires more than orchestrating containers — it demands control over the underlying hardware, especially GPUs, CPUs, and storage spread across different sites. Kubernetes provides logical isolation, but it doesn't manage physical resource allocation, security boundaries, or low-level performance tuning.

In practice, several key questions arise. How do you secure and monitor GPU access across multiple tenants and locations? Can your orchestration layer manage both virtual machines and containers while still working smoothly with NVIDIA technologies such as NVLink, InfiniBand, or DPUs? What about metering, usage tracking, or billing, which are essential for offering GPU or AI services internally or externally? And if you want to mix different business models — from GPU-as-a-Service to AI-as-a-Service or shared datasets

between teams — can your platform handle all that within one controlled environment?

There's a reason why even providers like AWS keep their infrastructure layer separate from their Kubernetes layer: both are important, but they solve different problems. In an AI Factory, both execution models coexist and complement each other:

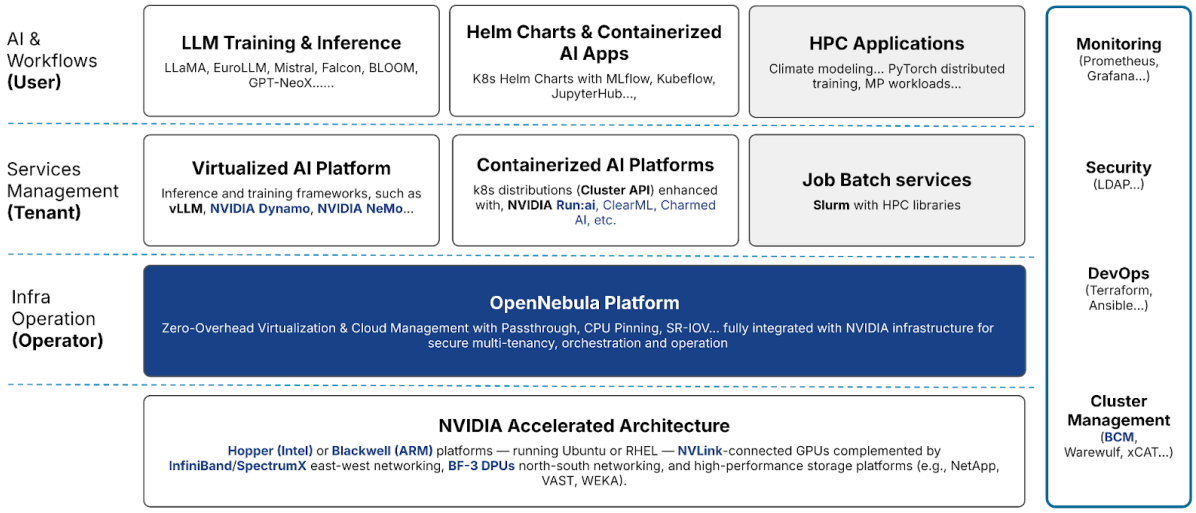
- **Cloud-Native AI** (Kubernetes-based): Best suited for multi-user setups, collaborative pipelines, and environments where developers need to manage and schedule multiple jobs efficiently. It offers flexibility and scalability, especially for model training and deployment workflows.
- **HPC AI** (Direct GPU Passthrough): The preferred approach when workloads are performance-sensitive or need tight control over hardware. Running directly on GPU-passthrough virtual machines gives near bare-metal performance, consistent latency, and predictable behavior — ideal for production inference or large, tightly managed training runs.

By combining both approaches under a single orchestration layer, organizations can balance efficiency, performance, and sovereignty, ensuring their AI workloads run where and how they want — without compromising speed or control.

3. Key Building Blocks

OpenNebula is the only virtualization and cloud platform that offers:

- **Full NVIDIA Hardware Integration**. Comprehensive support for the entire NVIDIA computing and networking ecosystem, across both x86 and ARM architectures and the latest Hopper and Blackwell GPU series. OpenNebula integrates MIG and GPU passthrough modes, NVLink, SpectrumX, InfiniBand, and DPUs within a single, unified orchestration layer.
- **Unified Orchestration of the NVIDIA Software Stack**. Delivers seamless management of NVIDIA components such as Run:ai, NIM, NeMo, and Dynamo under one orchestration framework—simplifying operations and ensuring consistent governance across environments.
- **Cloud-Like Access Model**. Delivers interactive, on-demand access for AI inference workloads, allowing users to deploy and scale resources instantly through a frictionless, self-service interface. Integrates seamlessly with Hugging Face models, enabling rapid deployment of state-of-the-art AI capabilities within a private, sovereign environment.
- **Flexible Execution Models**: Supports both HPC-style direct GPU passthrough for performance-critical inference and training, and Kubernetes-based cloud-native execution for multi-user pipelines and collaborative AI development.
- **Secure Multi-Tenancy**. Implements physically secure multi-tenancy through Virtual Data Centers (VDCs), offering guaranteed isolation, compliance, and predictable performance for different teams, customers, or projects.
- **Sovereign, Flexible, and Extensible Management Platform**. Built on open-source technology, OpenNebula provides a sovereign, vendor-neutral management layer that is fully extensible and future-proof—ready to integrate with emerging digital infrastructure and third-party tools.



OpenNebula integrates with several key components across the infrastructure ecosystem to build a complete AI Factory.

Tenant	Description	Examples
AI Platforms	AI development and execution platforms used by teams and departments to design, manage, and monitor AI workloads, as well as to deploy model endpoints for end users. Most of these platforms extend Kubernetes to optimize execution, monitoring, and scaling of machine learning and deep learning workloads across hybrid and edge environments.	vLLM, NVIDIA Dynamo, ClearML, NVIDIA Run:ai, Kubeflow, Charmed AI, Rancher AI
Kubernetes	Kubernetes serves as the orchestration backbone for many AI platforms, providing container scheduling, and automated scaling. In the context of an AI Factory, Kubernetes acts as a higher-level abstraction layer that manages AI frameworks and workloads within each tenant's Virtual Data Center (VDC), while OpenNebula provides the underlying infrastructure orchestration, GPU resource allocation, and secure multi-tenancy needed to run those environments efficiently.	Rancher RKE2, Red Hat OpenShift, Canonical Charmed Kubernetes
Storage	High-performance distributed or parallel file systems providing low-latency access to large datasets and supporting concurrent I/O from multiple GPU servers. Essential for training, inference, and large-scale data management in AI workloads.	NetApp, WekaFS, VAST Data
Monitoring	Observability and performance monitoring tools that collect and visualize system and application metrics to ensure availability, performance, and capacity optimization.	Prometheus, Grafana
Security	Identity and access management systems that enable centralized user authentication, authorization, and policy enforcement across distributed environments.	LDAP, Active Directory
DevOps	Automation tools that streamline infrastructure provisioning, configuration management, and continuous deployment workflows across AI Factory environments.	Terraform, Ansible
Cluster Management	Tools for provisioning and maintaining large-scale, bare-metal or virtual clusters, including node discovery, bootstrapping, and lifecycle management.	BCM, MAAS, Warewulf, xCAT

4. Secure Multi-tenant Model

In general, an AI Factory includes four distinct layers of users. In smaller deployments, some of these layers may be merged, while larger or public AI Factories might expose interfaces at all levels to support different business models.

Tenant	Description
End Users	These are the **consumers of AI services** who interact with inference models through a web portal or API endpoint. They typically authenticate via tokens and use standard interfaces such as OpenAI-compatible APIs. Their interaction is limited to consuming AI services rather than managing infrastructure or applications.
Service Users	Service Users are not OpenNebula users but rather **AI Platform users** who create or manage AI jobs through the APIs or web portals of each specific AI platform (e.g., for training, inference, or workflow execution). They operate within the services deployed and managed by the Tenant (Service Administrator), running AI workflows or pipelines without requiring any direct access to the underlying infrastructure.
Services Administrators (Tenants)	Tenants are **OpenNebula infrastructure users** who **manage and operate virtualized AI Platform** within assigned Virtual Data Centers (VDCs). They deploy and manage virtual machines or elastic workflows of VMs that encapsulate specific AI platform services such as inference engines (e.g., vLLM) or AI frameworks (e.g., NVIDIA Run:ai, NVIDIA Dynamo, or ClearML). Service Administrators are advanced users responsible for: • Deploying, configuring, and managing their service environments. • Ensuring service availability, performance, and security. • Integrating their service stack with the underlying infrastructure.
Infrastructure Administrators	A dedicated administration group within the AI Factory provider responsible for the operation of the physical and virtual infrastructure. They manage hardware, networking, storage, and the core OpenNebula platform, ensuring reliability, scalability, and security. **Infrastructure Administrators** define and allocate Virtual Data Centers (VDCs) that dynamically partition and isolate resources—clusters, hosts, datastores, and networks—among Tenants, enabling secure multi-tenancy and resource governance.

OpenNebula offers robust multi-tenancy capabilities through its **Virtual Data Center** (VDC)[1] model, enabling the logical segregation of resources—compute, storage, and networking—among multiple tenants. **ACLs (Access Control Lists)**[2] allow fine-grained assignment of permissions to different user roles, from infrastructure administrators to service-specific administrators and end users. This model supports both dedicated and shared resource allocation, ensuring security, isolation, and compliance while allowing each tenant to manage their own workloads independently within the shared infrastructure.

It also includes **accounting** and **billing** features that can be integrated with external **monetization or chargeback systems,** enabling detailed usage tracking and cost allocation across tenants.

A **common marketplace catalog** includes all the necessary images, templates, and multi-VM workflows for service platforms, enabling Service Administrators to deploy, operate, and expand services efficiently.

[1] https://docs.opennebula.io/stable/product/cloud_system_administration/multitenancy/manage_vdcs/
[2] https://docs.opennebula.io/7.0/product/cloud_system_administration/multitenancy/chmod/

5. Near-Zero Performance Overhead

Through the use of advanced passthrough technologies, OpenNebula ensures that running AI workloads, HPC jobs or Kubernetes clusters inside virtual machines (VMs) introduces no measurable performance degradation. When configured with PCI **passthrough**, SR-IOV, and CPU pinning, VMs and containers have direct access to underlying hardware resources, achieving performance equivalent to bare metal.

By default, nodes are dedicated to a single VM to guarantee maximum throughput; however, when flexibility is required, OpenNebula supports resource-sharing configurations that maintain near-native performance. In this proposal, passthrough is specifically applied to both the **high-performance communication network and the storage network**, ensuring that the solution operates without any performance penalty.

Several studies[3] have shown that the **performance overhead introduced by virtualization in AI workloads is minimal**—ranging from slightly below 0% to a maximum of about 5% for both single-VM and distributed training/inference scenarios, even when using network virtualization. Interestingly, cases of "negative overhead" (<0%) have been reported, meaning that virtual machines can actually run *faster* than bare metal. This typically happens because VM environments are often configured with performance optimizations by default—such as large pages (hugepages) and NUMA-aware scheduling—whereas bare-metal systems are not always tuned in the same way.

More than ten years ago, in 2012, FermiLab published performance results[4] from running an OpenNebula-based cloud infrastructure over InfiniBand, **demonstrating that virtualization technology could achieve nearly 100% efficiency** in both bandwidth and latency benchmarks. This early validation confirmed OpenNebula's ability to combine cloud flexibility with HPC-grade performance, a capability that now underpins its AI Factory architecture—enabling modern enterprises to run GPU-accelerated, low-latency AI workloads with the same near-native efficiency once proven in high-performance computing environments.

6. Conclusions and Next Steps

The OpenNebula AI Factory Reference Architecture offers a proven framework for building sovereign, high-performance, and scalable AI infrastructure. It is particularly well-suited for organizations seeking to:

- Deliver AI-as-a-Service or GPU-as-a-Service offerings with full control and cost predictability.
- Consolidate AI, HPC, and edge workloads within a single, secure, multi-tenant environment.
- Ensure data sovereignty and vendor neutrality through an open, extensible, and EU-based management platform.

At OpenNebula, we have extensive experience deploying AI Factories for a diverse range of customers — from **cloud providers** and **telecommunications** operators to **HPC centers** and enterprise research organizations. While the core platform capabilities remain consistent, each deployment is tailored to the specific operational needs of the customer: HPC users typically integrate OpenNebula with batch job schedulers like SLURM or xCAT, while Telco environments leverage it to connect with 5G+ edge infrastructure for AI-driven, low-latency services.At

Contact us to schedule a live **demonstration** of the OpenNebula AI Factory in action. Our team will guide you through a Proof of Concept (**PoC**) and **pilot** deployment, helping you design a future-ready AI infrastructure that combines cloud agility, HPC performance, and sovereign control.

[3] https://research.ibm.com/publications/to-virtualize-or-not-to-virtualize-ai-infrastructure-a-perspective
[4] https://lss.fnal.gov/archive/2014/conf/fermilab-conf-14-470-cd.pdf

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_20251107