Open Nebula

GUIDE

AI Factory Integrated Platform Support

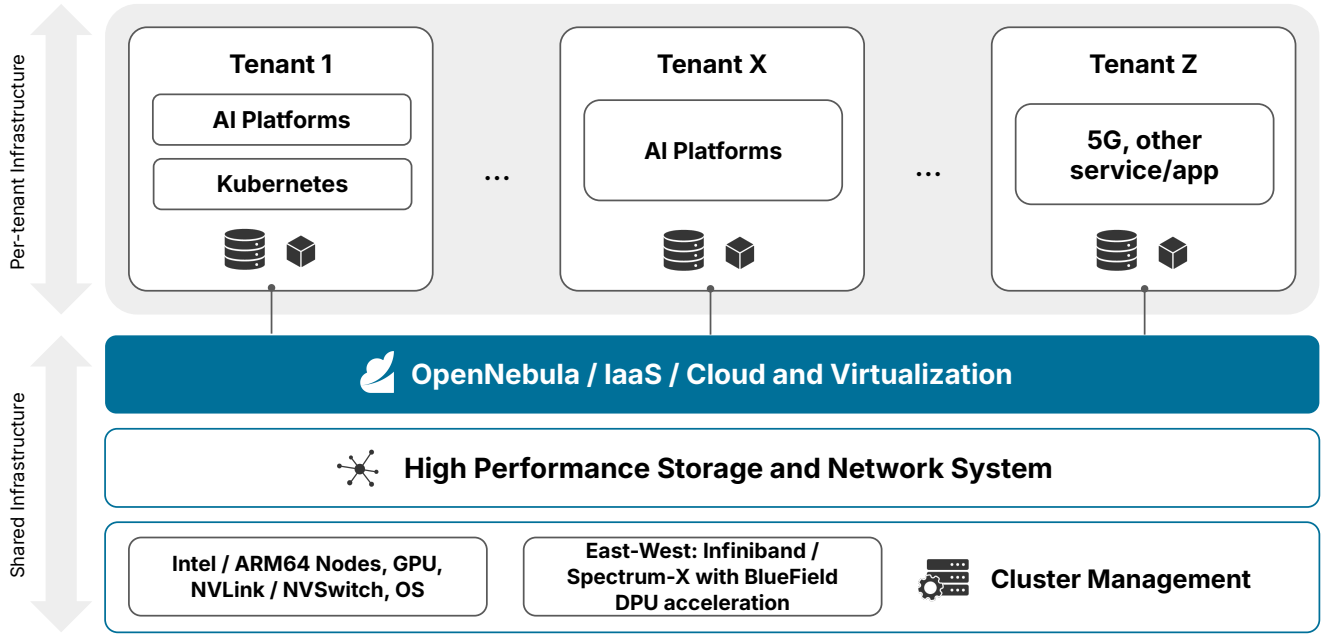
The OpenNebula AI Factory Support Add-on extends the standard OpenNebula Subscription by providing specialized assistance for deploying and operating AI workloads on OpenNebula-based infrastructures. Rather than introducing new software components, this add-on focuses on helping organizations enable, optimize, and manage GPU-accelerated environments for AI and machine learning applications, in full alignment with the OpenNebula AI Factory Reference Architecture.

The Add-on offers expert guidance on GPU orchestration, integration with AI frameworks and pipelines for inference, fine-tuning, and training, as well as performance optimization and best practices for running large-scale workloads. By leveraging OpenNebula's native virtualization and orchestration capabilities—combined with technologies such as Prometheus, Grafana, InfiniBand, NVLink/NVSwitch, and HPC/AI storage backends—this add-on ensures that customers can fully exploit the power of their AI infrastructure.

AI Factory Reference Architecture

The OpenNebula AI Factory Support Add-on provides the expert guidance required to deploy and operate infrastructures aligned with the OpenNebula AI Factory Reference Architecture. This architecture defines a clear separation between the shared infrastructure layer, managed centrally, and the tenant-level environments, where individual teams operate independently within their own isolated Virtual Data Centers (VDCs). The shared layer integrates high-performance GPU servers (such as NVIDIA H100, GB200, and GB300) interconnected through NVLink and high-speed fabrics like InfiniBand or Spectrum-X with BlueField-3 DPUs, as well as distributed storage systems that ensure efficient data access across workloads and tenants.

The add-on helps customers configure, validate, and optimize this architecture—covering GPU orchestration, storage and network integration, and virtualization through the IaaS layer. It also supports the deployment of tenant environments that run AI platforms either on Kubernetes for multi-user scheduling and pipelines or directly on GPU servers for performance-critical workloads. By aligning with the AI Factory Reference Architecture, the add-on ensures that organizations can build secure, scalable, and high-performance AI infrastructures on OpenNebula with expert technical assistance throughout the process.



What the Support Add-on Includes

The OpenNebula AI Factory Support Add-on provides a specialized support service that extends the standard Subscription to cover AI and GPU-intensive environments. This add-on ensures that organizations can deploy and operate AI workloads confidently—focusing on the performance, and efficiency of the infrastructure and orchestration layers, rather than the AI frameworks themselves.

- ✔ **AI Workload Enablement:** Deploying and tuning AI frameworks within OpenNebula environments—for HPC-AI workloads (e.g., vLLM) and container-native deployments (e.g., NVIDIA Dynamo, Run:ai).
- ✔ **Infrastructure and Orchestration:** Support for GPU orchestration, storage configuration, advanced networking (InfiniBand, Spectrum-X, NVLink), automation, VM lifecycle, and container orchestration via Kubernetes (Cluster API).
- ✔ **Integration and Optimization:** Assistance with aligning virtualization, networking, GPU, and storage systems for efficient AI execution in line with the OpenNebula AI Factory Architecture.
- ✔ **Operational Best Practices:** Expert recommendations for performance optimization, capacity planning, and secure multi-tenant AI environments.

What the Support Add-on Excludes

The AI Factory Support Add-on focuses on assisting with infrastructure configuration, integration, and operations rather than on third-party software or AI model development. The following points clarify the boundaries of this service:

- ✔ **Third-Party Certified Components:** Support applies to the integrated environment as a whole, not to independent third-party software or frameworks. Coverage is limited to certified versions in release notes and deployed using AI Factory–compliant architectures and official procedures.
- ✔ **AI Frameworks and Model Code:** Issues related to AI frameworks, libraries, or model fall under the responsibility of their respective providers or open-source communities. Support focuses on the infrastructure and orchestration layers, including deployment, and operational guidance.
- ✔ **Upstream Dependencies:** Problems identified in upstream components (e.g., kernel, libraries, or external packages) will be escalated to the relevant projects.
- ✔ **Immediate Fixes or Enhancements:** The add-on does not include on-demand fixes, patches, or feature enhancements for third-party software.
- ✔ **Operational Responsibility:** Customers are expected to have qualified personnel capable of operating AI Factory environments and validating new frameworks or models.

The scope of Severity 1 and 2 incidents is limited to issues that impact the OpenNebula core platform or its supported infrastructure components. Problems involving third-party or community software (such as AI frameworks, libraries, or drivers) are handled on a best-effort basis and may be reclassified as Severity 3 or 4, depending on their operational impact.

Pricing

All support services under the AI Factory Support Add-on are provided as an extension of the standard OpenNebula Subscription, with an annual fee based on the number of GPUs covered.

OpenNebula Systems USA

1500 District Ave
Burlington, MA 01803, USA

OpenNebula Systems Europe

Paseo del Club Deportivo 1 – Edificio 4 Planta 1
Parque Empresarial La Finca
28223 Pozuelo de Alarcón, Madrid, Spain

Reference: AI Factory Integrated Platform Support - Rev20251215