



# OneAI

| An Open-Source Framework for Managing AI Models at Scale

# Who am I?

Daniele Mingolla - Software Developer for AI at OpenNebula Systems



- **Background**
  - MSc in Data Science (University of Milano-Bicocca)
  - BSc in Computer Science (University of Perugia)
- **Previous experience**
  - Data Scientist in mobile gaming, logistics, and food delivery
  - Worked on ML systems, A/B testing, and data engineering
- **Current focus**
  - Implementing AI features as part of the Innovation team in OpenNebula

# Who is OpenNebula?

The Open Source Cloud & Edge Computing Platform



- First open source **laaS** solution, created **17 years ago**, with a **vibrant user community**.
- **Enterprise infrastructure software** company with **15 years of experience**.
- HQs in **Madrid (Spain, EU)** and **Burlington (MA, US)**, and offices in **Brussels (BE, EU)** and **Brno (CZ, EU)**.
- More than **5,000 clouds** worldwide, largest with **16 DCs** and **300K cores**.



# Table of Contents

1

## OpenNebula

What Is OpenNebula, AI Factories

2

## Why OneAI

The Problem, The Solution

3

## Architecture

High-Level Overview

4

## The Three Pillars

Marketplace, Datastore, Inference API

5

## From Command to API

CLI Flow, GPU Scheduling

6

## Wrap-Up

Demo, Summary & Benefits

1

# OpenNebula

# What is OpenNebula?

The Open Cloud & Edge Computing Platform

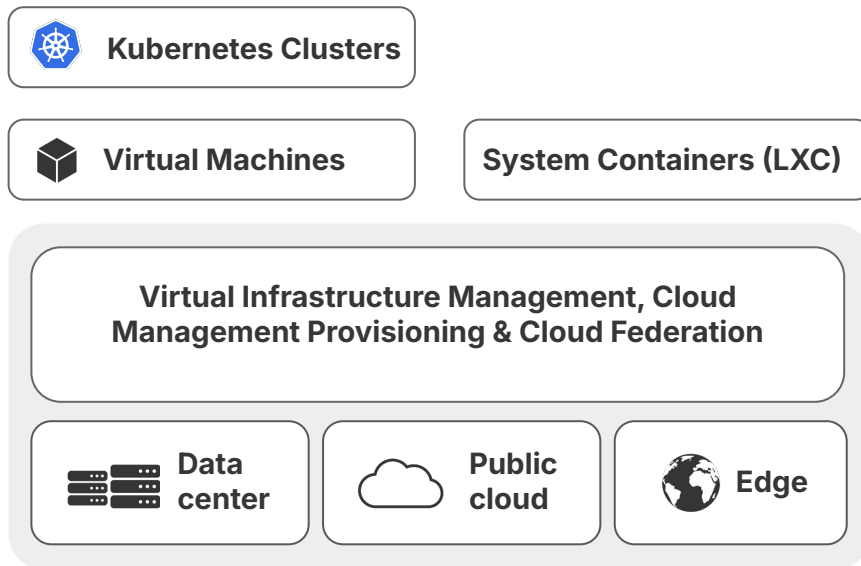
## VIM Functionality

- Simplicity and light profile
- Extensible architecture
- Multi-tenancy & Multi-VM
- DevOps friendly

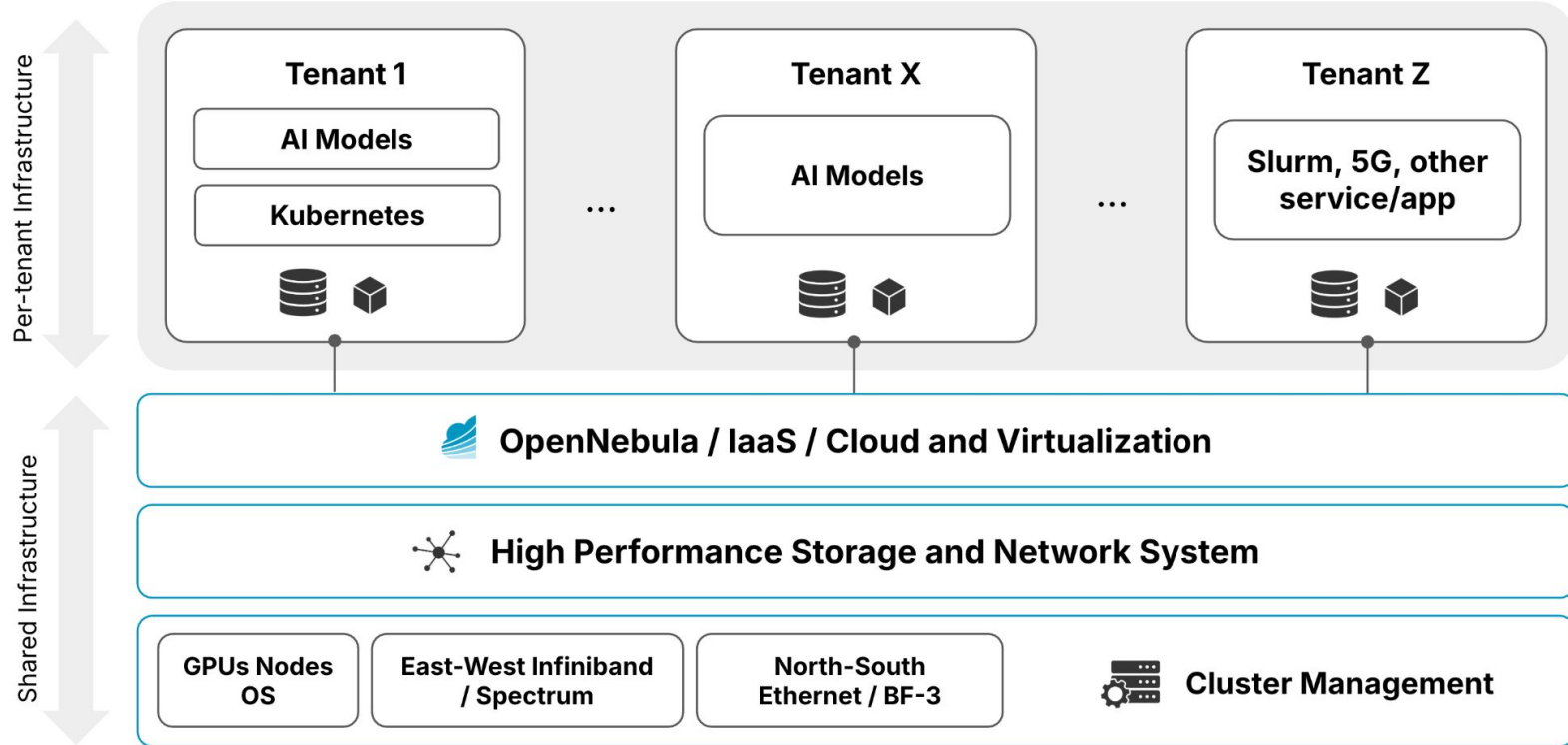
## OneKE (Kubernetes)

- Virtual appliances KaaS
- Different "add-ons"

## Cloud-Edge Continuum Apps



# AI Factories



2

## Why OneAI

# The Problem

Running AI inference at scale is fragmented

## 1. Model discovery

Where are my models? How do I find them?

## 2. Storage

How do I store 100GB+ model files efficiently?

## 3. Deployment

How do I get a model running on GPUs?

## 4. API exposure

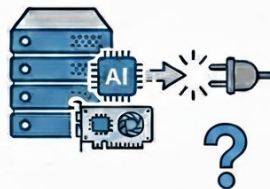
How do I serve it to applications?



AI MODEL DISCOVERY



STORAGE



DEPLOYMENT GAP



API BARRIER

# OneAI: The Solution

Open-source framework to run AI models and inference on OpenNebula

## 1. Hugging Face Hub Marketplace

Lightweight catalog; metadata-first

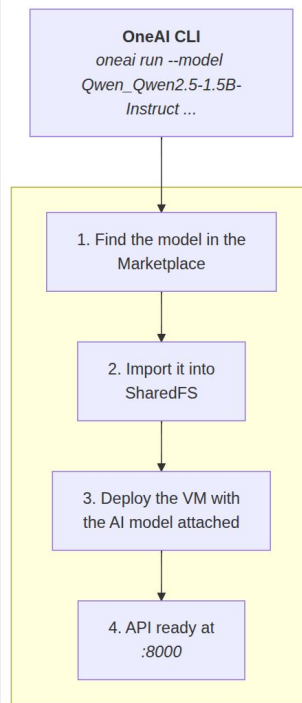
## 2. SharedFS Datastore

Directories as images; high-performance shared storage

## 3. vLLM + OpenAI API

Orchestration + standard endpoints

Reuse **high-performance storage** and supports secure **multi-tenancy**

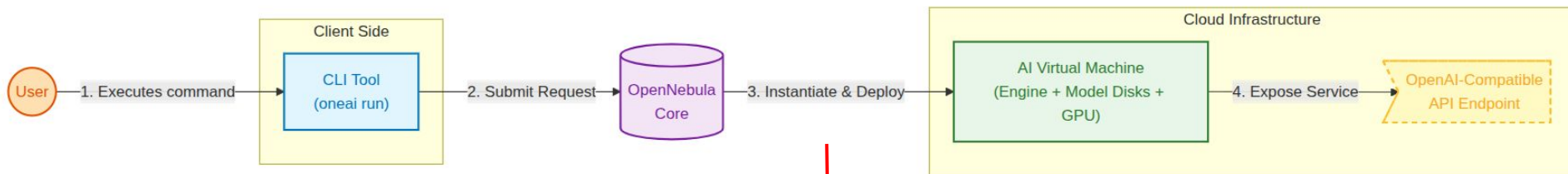


3

# Architecture

# High-Level Architecture

**No extra orchestration service:** you run the **CLI**, it talks to **OpenNebula**, and OpenNebula schedules and runs the VM.



```
root@localhost:~# curl http://10.0.1.184:8000/v1/completions -H "Content-Type: application/json" -d '{
  "prompt": "What is OpenNebula?",
  "max_tokens": 50,
  "temperature": 0
}' | jq '.choices[0].text'
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total     Spent    Left  Speed
100  772  100    701  100    71   2877    291  --:--:--  --:--:--  --:--:--  3176
" OpenNebula is an open-source virtualization platform that allows users to create, manage and scale cloud-based computing resources. It provides a web-based interface for managing virtual machines (VMs) and storage pools, as well as APIs for integration with"
```

4

# The Three Pillars

# Pillar 1: Hugging Face Hub Marketplace

Lightweight catalog of Hugging Face models inside OpenNebula marketplace

## Metadata-only catalog

Models are listed in the OpenNebula marketplace as apps (metadata). The actual model files are not stored there yet.

## Query the AI model by name

You call the CLI with a model name. The CLI looks up that name in the marketplace and resolves it to an image.

## Materialize on deploy

Artifacts are copied only when you deploy. The catalog points to Hugging Face; the image is created in your datastore only when needed.

Attributes	
DESCRIPTION	HuggingFace model: Qwen/Qwen2.5-14B-Instruct
DOWNLOADS	3617202
HAS_CHAT_TEMPLATE	YES
IMPORT_ID	-1
LICENSE	unknown
LINK	<a href="https://huggingface.co/Qwen/Qwen2.5-14B-Instruct">https://huggingface.co/Qwen/Qwen2.5-14B-Instruct</a>
METADATA_ONLY	YES
MODEL_ID	Qwen/Qwen2.5-14B-Instruct
PUBLISHER	huggingface.co
TASK	text-generation
VERSION	1.0

# Pillar 2: SharedFS Datastore

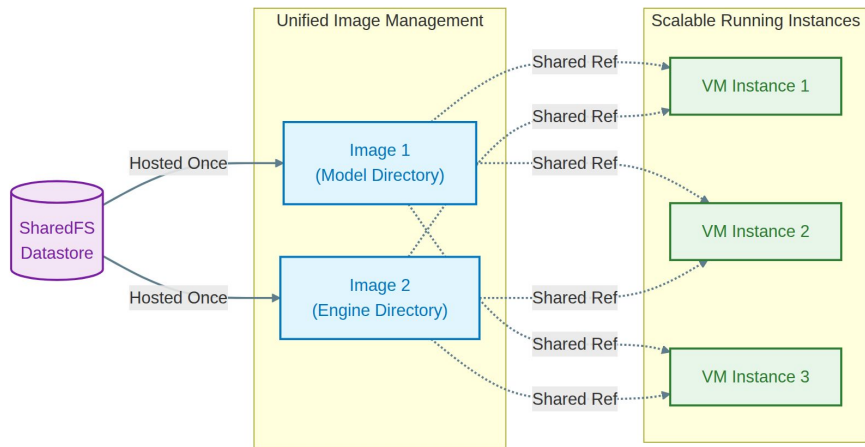
Datastore where directories on shared storage count as VM images (no copy)

## Why shared storage

Large model files (10GB–100GB+) are expensive to copy. Shared high-performance storage is already in place for HPC workloads. Keeping models there avoids duplication and transfer overhead.

## Benefit

Multiple VMs can mount the same model directory simultaneously. One copy serves many deployments, reducing storage costs and speeding up provisioning.



# Pillar 3: Inference VM + OpenAI API

Deploys vLLM with your model and exposes an OpenAI-compatible API

## VM provisioning

OneAI generates a template (model + engine disks, GPU, config); OpenNebula creates the VM, assigns a GPU, and starts it.

## Inference engine

Inside the VM, the vLLM appliance loads the model from disk. Configuration options are passed via context variables when the VM starts.

## OpenAI-compatible API

Each VM exposes an endpoint that speaks the OpenAI API. Any application that works with OpenAI API can use it directly, no code changes needed.

```
root@vgpu03: ~  
oneadmin@vgpu03:/root$ curl http://10.0.1.184:8000/v1/completions -H "Content-Type: application/json" -d '{"prompt": "What is OpenNebula?", "max_tokens": 50, "temperature": 0}' | jq '.choices[0].text'  
% Total % Received % Xferd Average Speed Time Time Time  
Current                                Dload Upload Total Spent Left  
Speed  
0 0 0 0 0 0 0 0 --:--:-- --:--:-- --:--:--  
100 772 100 701 100 71 3993 404 --:--:-- --:--:-- --:--:--  
-- 4411  
" OpenNebula is an open-source virtualization platform that allows users to create, manage and scale cloud-based computing resources. It provides a web-based interface for managing virtual machines (VMs) and storage pools, as well as APIs for integration with"
```

5

# From Command to API

# From CLI to running Inference API

```
oneai run --model Qwen_Qwen2.5-1.5B-Instruct --engine-image-id 17 --network  
SERVICE --cpu 8 --memory 10Gi --gpu 1 --gpu-type h1001
```



```
root@vgpu03: ~  
oneadmin@vgpu03:/root$ curl http://10.0.1.184:8000/v1/completions -H "  
Content-Type: application/json" -d '{  
  "prompt": "What is OpenNebula?",  
  "max_tokens": 50,  
  "temperature": 0  
}' | jq '.choices[0].text'  
% Total    % Received % Xferd  Average Speed   Time    Time     Time  
Current                                  Dload  Upload  Total  Spent  Left  
  
Speed  
0      0    0    0    0    0      0      0  --:--:--  --:--:--  --:--:--  
100  772  100  701  100  71    3993    404  --:--:--  --:--:--  --:--:--  
--  4411  
" OpenNebula is an open-source virtualization platform that allows use  
rs to create, manage and scale cloud-based computing resources. It pro  
vides a web-based interface for managing virtual machines (VMs) and st  
orage pools, as well as APIs for integration with"
```

# GPU Scheduling

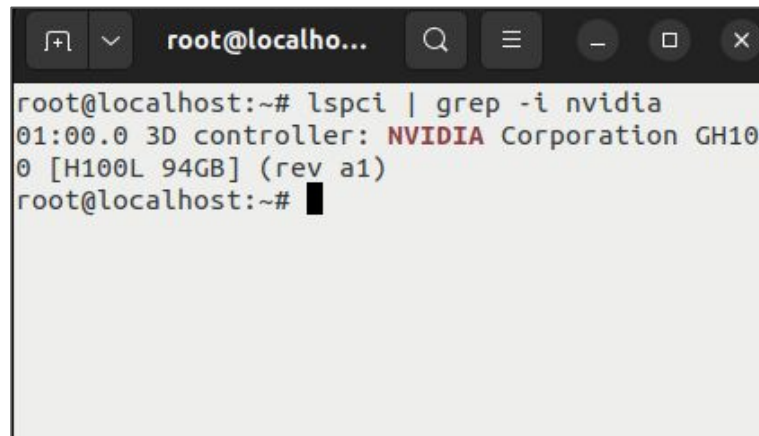
Deploys vLLM with your model and exposes an OpenAI-compatible API

## GPU matching

You pass GPU count and type in the CLI. OpenNebula scheduler matches it to a free GPU on a host with PCI pass-through configured. If you omit GPU, only CPU is used.

## GPU pass-through to VM

The host passes the GPU through to the VM via PCI pass-through. The VM sees the GPU directly and vLLM can use it for inference. Infrastructure-native placement; **no custom scheduler needed.**



```
root@localho...
root@localhost:~# lspci | grep -i nvidia
01:00.0 3D controller: NVIDIA Corporation GH10
0 [H100L 94GB] (rev a1)
root@localhost:~#
```



# Demo

6

## Wrap-Up

**OneAI** lets you **discover a model** from Hugging Face, **store it** in a shared datastore, **deploy an inference engine** with that model & **expose an OpenAI-compatible API**.

- **Reproducible:** Same command, same result. Adapts to automation workflows (scripting, MCP tools).
- **Your control:** Your models, your infra. OpenNebula and standard OpenAI API.
- **Reuse storage:** One copy of an AI model serves many VMs. Lower storage costs.
- **Scale:** OpenNebula handles isolation and resources across the cluster.

**OneAI** and the **SharedFS datastore** are **currently in development** and will be available in a future OpenNebula release.

# Thank You!

# Discover OpenNebula!

Build Your OpenNebula Cloud in 5 Minutes!

## Try the New OpenNebula 7.0 “Phoenix”



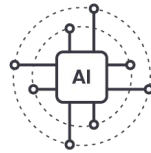
### VMware Revirtualization

AI-driven DRS, enterprise integrations (Veeam, NetApp), and efficient workload migration tools.



### Hybrid Multi-Cloud

On-demand deployment across multiple providers with seamless public cloud integration.



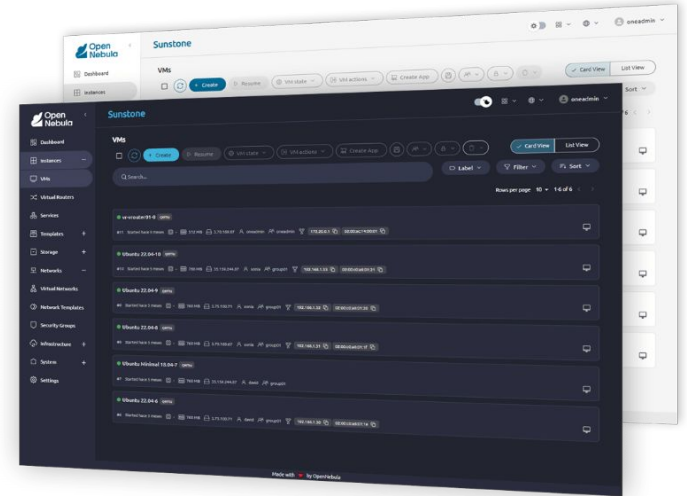
### AI Factories

GPU support with vLLM, Hugging Face, Ray, and NVIDIA frameworks for training and inference.



### Modernized Operations

Intelligent automation, predictive monitoring, and distributed infrastructure management.



<https://opennebula.io/opennebula-7>

## AI Plumbers :

OneAI: An Open-Source Framework for Managing AI Models at Scale.

## Network :

Building an Open Source Private 5G Network: A Practical Blueprint.

## Virtualization & Cloud Infrastructure :

How I Turned a Raspberry Pi into an Open-Source Edge Cloud with OpenNebula.

Arming Cloud Computing Continuum: Hunting vulnerabilities in open source hybrid clouds.



***Find our Booth in LEVEL 1 of BUILDING K***



ONEnextgen

> [OpenNebula.io/IPCEI-CIS](https://OpenNebula.io/IPCEI-CIS)

# IPCEI-CIS

## Next-Generation European Platform for the Datacenter-Cloud-Edge Continuum

Initiative supported by the Spanish Ministry for Digital Transformation and Civil Service through the **ONEnextgen Project: Next-Generation European Platform for the Datacenter-Cloud-Edge Continuum** (UNICO IPCEI-2023-003) and co-funded by the European Union's NextGenerationEU instrument through the Recovery and Resilience Facility (RRF).

