Building AI Factories with Open Source Tools

Aleksejs Petrovs (OpenNebula Systems)

IPCEI-CIS

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

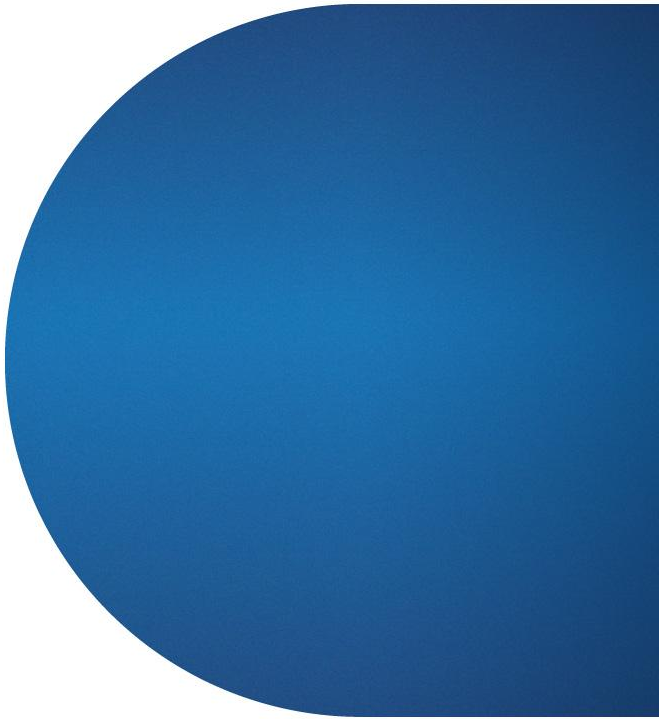
Initiative supported by the Spanish Ministry for Digital Transformation and Civil Service through the **ONEnextgen Project: Next-Generation European Platform for the Datacenter-Cloud-Edge Continuum** (UNICO IPCEI-2023-003) and co-funded by the European Union's NextGenerationEU instrument through the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

OpenNebula.io/IPCEI-CIS

Agenda

What Are We Going to Cover Today

- Problem & Concerns!
- How we can solve them?
- What is Ray appliance
- Demo



AI Factories

Let's Define the AI Factories Together!

AI Factory -

The Problem

What is the Problem?

No easy way to start with using LLMs on-premise or with the private cloud!

SaaS offerings might not suit your needs or too expensive to in a long run!

When you are going beyond "my computer" - it requires a lot of components and solutions to make it right!

Every public cloud vendor has its own way of configuration.

The Solution!

OpenNebula + Ray appliance

We brought the "AI as a Service" inference to your datacenter.

Can run your custom Python code.

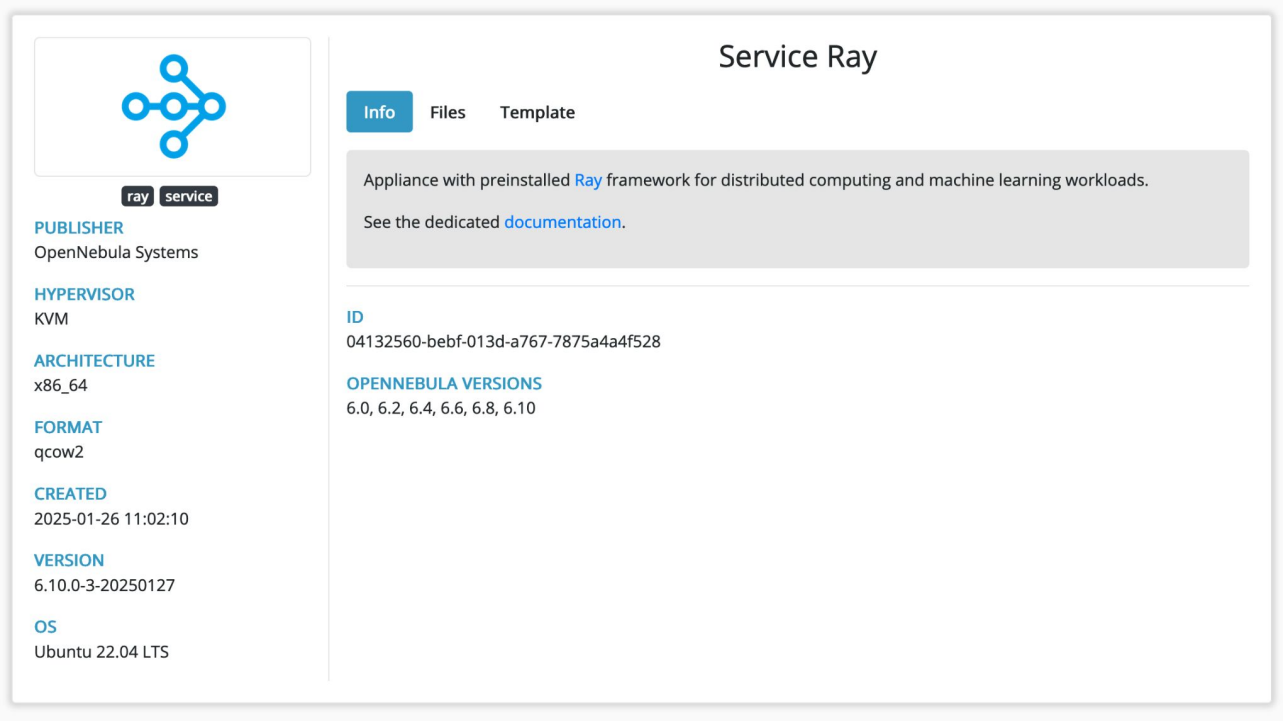
GPU passthrough and SR-IOV functionality for better performance!

A few-click deployment using the pre-built appliance to run your custom code and one of the certified LLMs from Hugging Face.

Aside from Ray - OpenNebula offers other advanced features of the private or hybrid cloud.

Service Ray

The Ready-to-Use Appliance



Tips on Deployment

Input Variables, Sizing & Fine-Tuning

- By default, Ray appliance comes tiny (disk size is 8GB). Scale it up so it can fit the selected LLM model.
- Supports the following Input Variables:
  - **ONEAPP_RAY_API_PORT** - Port to listen on
  - **ONEAPP_RAY_MODEL_ID** - The model name to download from HF
  - **ONEAPP_RAY_MODEL_TEMPERATURE** - Finetune the Temperature
  - **ONEAPP_RAY_MODEL_TOKEN** - HF API Key
- At least 8G of Memory is required for running the appliance.
- You can upload your own Python script to run inside the appliance using either URL or paste directly encoded in Base64.

Why OpenNebula for Enterprise AI?

Unlock **the Power of AI at the Edge** with
OpenNebula NextGen

ONEnextgen

Simplify LLM Deployment

An intuitive and simple platform for deploying and managing private clouds for LLMs.

Reduced Operational Costs

Cost-effective alternative to proprietary solutions like VMware, Nutanix or Red Hat or public cloud providers.

Native Support for GPUs

Out-of-the-box support for GPU virtualization, dynamic allocation and passthrough, ensuring optimal performance for AI and ML workloads.

Robust Multi-Tenancy

Users and Groups, Quotas and accounting, and VDC (virtual data-centers)

Unified Hybrid Cloud

Extend on-prem with public cloud clusters with uniform provisioning interface and operational procedures.

Deploy Hugging Face LLMs

Integrate validated LLMs for GenAI directly from Hugging Face to run on your VMs.

What's Next?

The Future of the Appliance

The following new features and improvements are currently being planned for the next release:

- Support for vLLMs
- OpenAI API
- Extended list of LLMs + recommended sizes

OpenNebula.io

contact@opennebula.io

+34 91 297 9741 / +1 781 238 6643

OpenNebula Systems Headquarters

EMEA

La Finca Business Park, Building 13
28223 Pozuelo de Alarcón, Madrid
Spain

USA

1500 District Avenue
Burlington, MA 01803
USA

OpenNebula Labs

Czech Republic

Cyrilská 7 – Impact Hub Brno
602 00 Brno
Czech Republic

Belgium

Brussels Manhattan Center, 5th Floor
Avenue du Boulevard 21, Brussels 1210
Belgium