



High Performance Infrastructure on OCI For AI and Beyond

Dr. Leo Li

AI Infrastructure Architect

JAPAC Center of Excellence (CoE)

Oracle CAPAC Services (Singapore Branch)

1, Fusionopolis Place, Level 12 Galaxis, Singapore 138522

Safe Harbor Statement

This presentation is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Statements in this presentation relating to Oracle's future plans, expectations, beliefs, intentions, and prospects are "forward-looking statements" and are subject to material risks and uncertainties. A detailed discussion of these factors and other risks that affect our business is contained in Oracle's Securities and Exchange Commission (SEC) filings, including our most recent reports on Form 10-K and Form 10-Q under the heading "Risk Factors." These filings are available on the SEC's website or on Oracle's website at <http://www.oracle.com/investor>. All information in this presentation is current as of January 2024 and Oracle undertakes no duty to update any statement in light of new information or future events.

Menu

- Why Oracle for AI?
 - Supercluster: the first principles behind OCI AI infrastructure
 - OCI AI stack and offering, prebuild and customization
 - OCI's partnership and integration for agentic AI
- Demos:
 - Deploy HPC and GPU cluster on OCI in minutes, *networking engineer easy!*
 - Deploy Kubernetes with RDMA and GPU on OCI in minutes, *devops easy!*
 - Deploy Nvidia NIM on OCI GPU clusters in minutes, *ai infra engineer easy!*

Why Oracle for AI?

The Problems of Morden AI/ML Workloads

Morden AI and ML based on GPT need massive computing power in

- Powerful processors
- Fast and dense local storage
- High-throughput and low-latency networking
- Cloud native and vector database for Agentic AI and beyond

Oracle's Approach: Enterprise-grade Cloud for AI/HPC

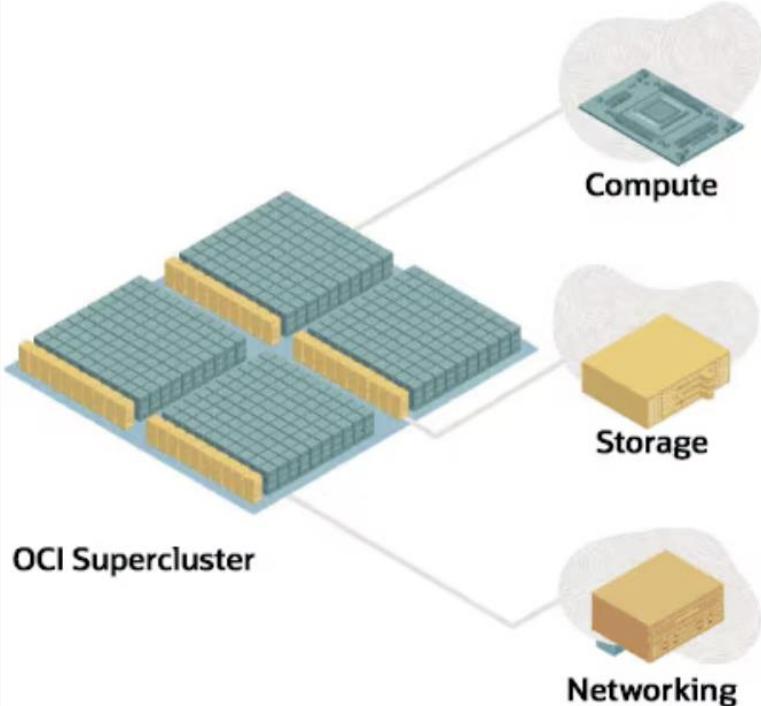
- On-premises-like performance and stability
 - Bare metal servers: no hypervisor, no virtualization penalty
 - **Real HPC RDMA network**: up to 3,200Gbps/server bandwidth + low μs latency
 - **NVMe on every node**
- With the flexibility, cost savings, and functionality of the cloud
- Partnership with Nvidia, OpenAI, Meta, xAI and more
- Managed opensource tools like **SLURM**, **Kubernetes** and more

OCI AI Infrastructure

The Supercluster Architecture powered OCI AI Infrastructure

OCI AI Infrastructure is the world leading high performance cloud based on ***high-speed nonblocking RDMA network fabric with unparallel super scalability in production***

Architecture



Overview

Supercharged GPU

- NVIDIA H200, H100, L40S, A100
- AMD MI300X
- DPU with RDMA

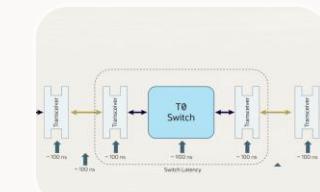
Massive capacity and high-throughput storage

- up to 61.44TB NVMe SSD
- 80 Gb/sec throughput

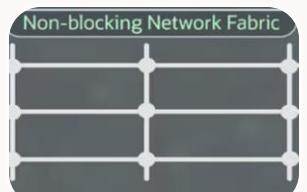
Ultrafast nonblocking network

- 2.5-9.1 us of network latency (3 layer)
- up to 3.2 Tb/s network bandwidth
- up to 200 Gb/sec of front-end bandwidth

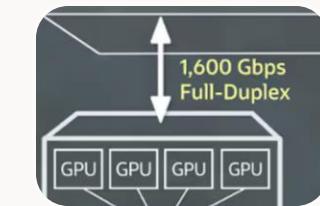
Features



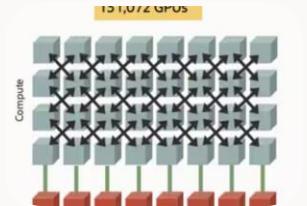
2.5 us latency



nonblocking RDMA

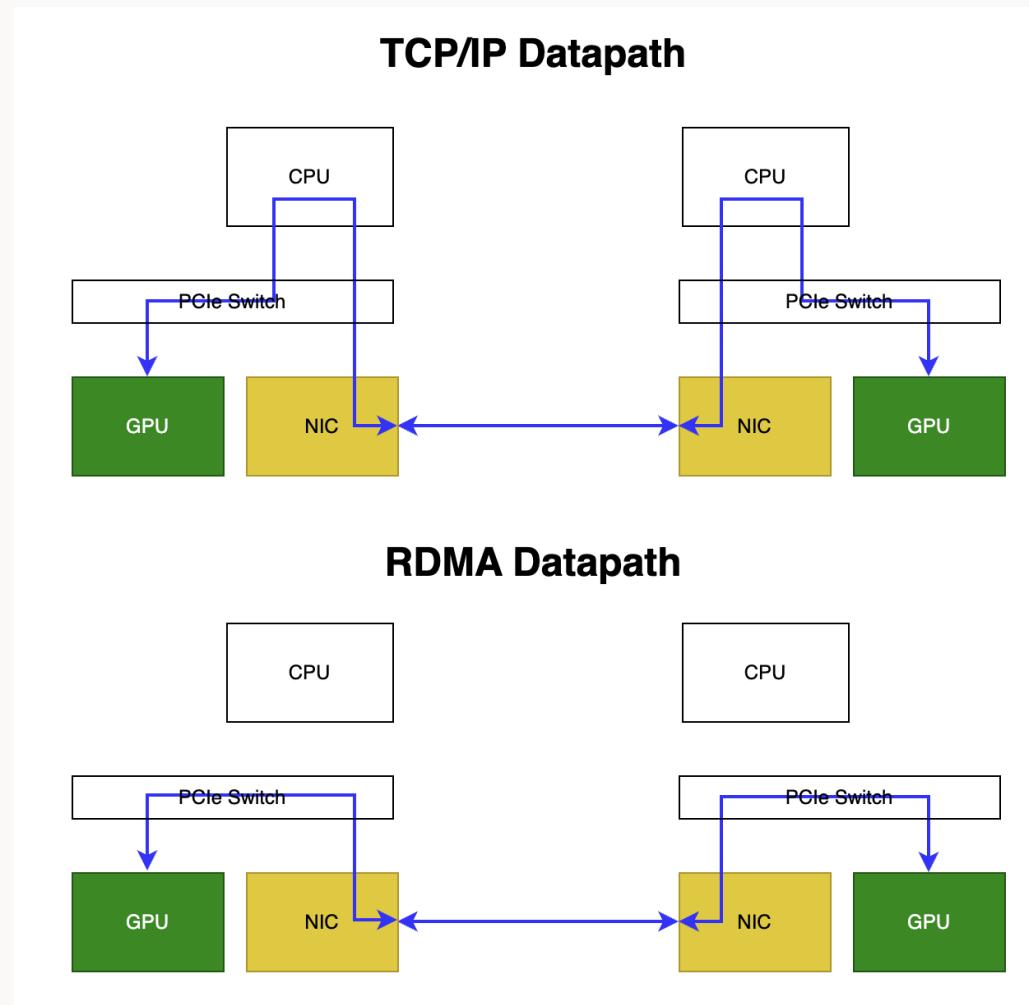


3.2Tb/s Bandwidth



131,072 GPUs Scalability

Why RDMA? Bypass CPU, Low Latency and Full Bandwidth

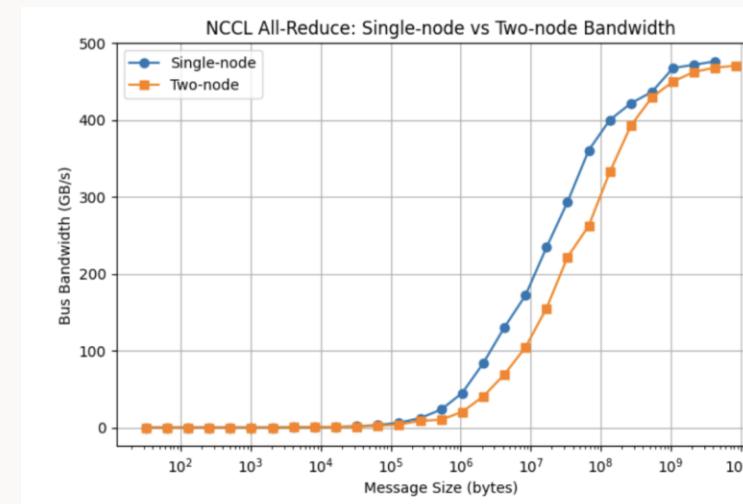
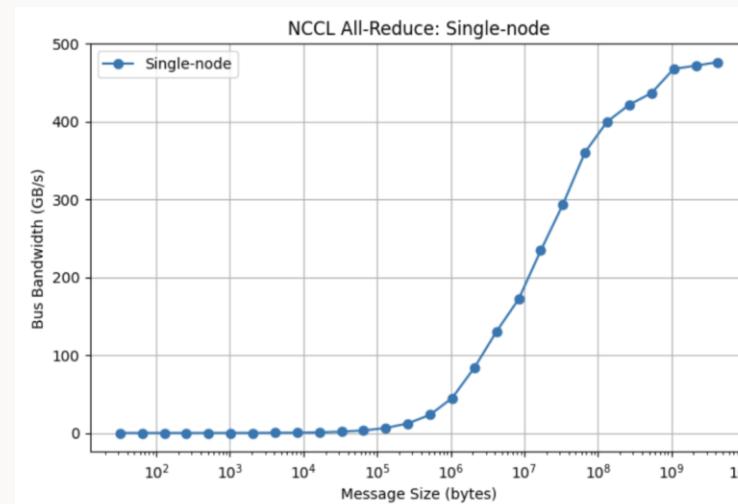
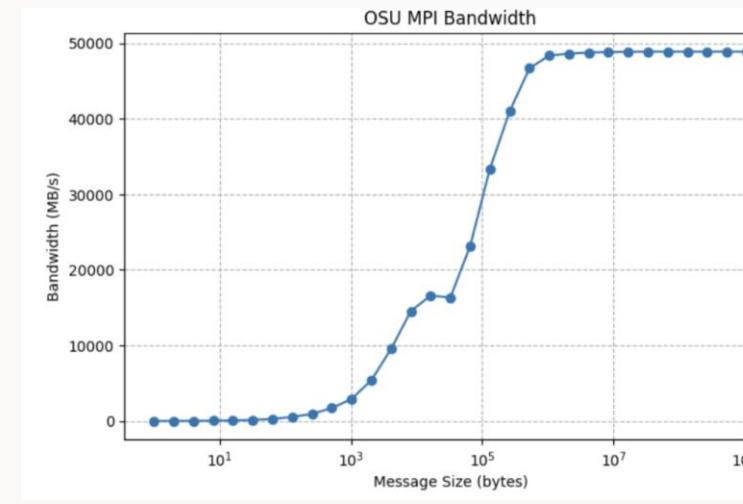
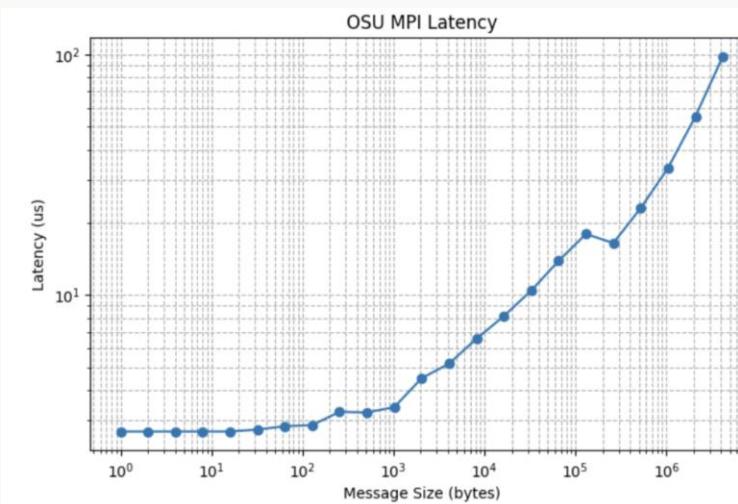


RDMA vs TCP/IP

Features	Benefit
Bypass CPU	lower CPU usage
No OS context switch	lower latency
Shorter PCIe data path	lower latency
PCIe Switch Downlink P2P	full bandwidth*

*In multi-GPU systems with PCIe switches, traditional TCP/IP networking often relies on CPU-mediated paths and PCIe uplink bandwidth. This introduces bottlenecks that constrain data throughput and increase latency — especially under load. In contrast, RDMA—particularly with GPUDirect—enables direct memory-to-memory transfers between NICs and GPUs, bypassing the CPU and reducing pressure on PCIe uplinks. This architecture is essential in large-scale GPU clusters, where high-frequency, low-latency communication directly impacts training efficiency.

Metrics Talk: the State-of-the-Art Performance of OCI Supercluster



OSU MPI Benchmarking

- RDMA benefit
- <3us latency, worst ~ 9us in larger scale
- 97% throughput

NCCL MPI Benchmarking

- RDMA benefit
- 97% Scaling of Collective Operations

OCI AI Infrastructure Offerings



With the **innovation of isolation technology** on RDMA network, OCI AI infra offers the scalability of **every size** to our customers' need. With **premium better pricing** compared to competitors.

OCI AI Infrastructure Products

Oracle Roving Edge Infrastructure



8 – 48 OCPUs

NVIDIA L4 GPUs

OCI IaaS, OKE, select OCI Services, Oracle & 3rd party marketplace solutions

Oracle Compute Cloud@Customer



552 – 6600+ OCPUs

NVIDIA L40S GPUs

OCI APIs, CLI, SDKs, Tools, and User Experience

OCI Dedicated Region



Full Spectrum of OCI Region IaaS

NVIDIA H100, A100, L40S, A10, V100, AMD MI300X

150 + OCI Services

OCI Public Region

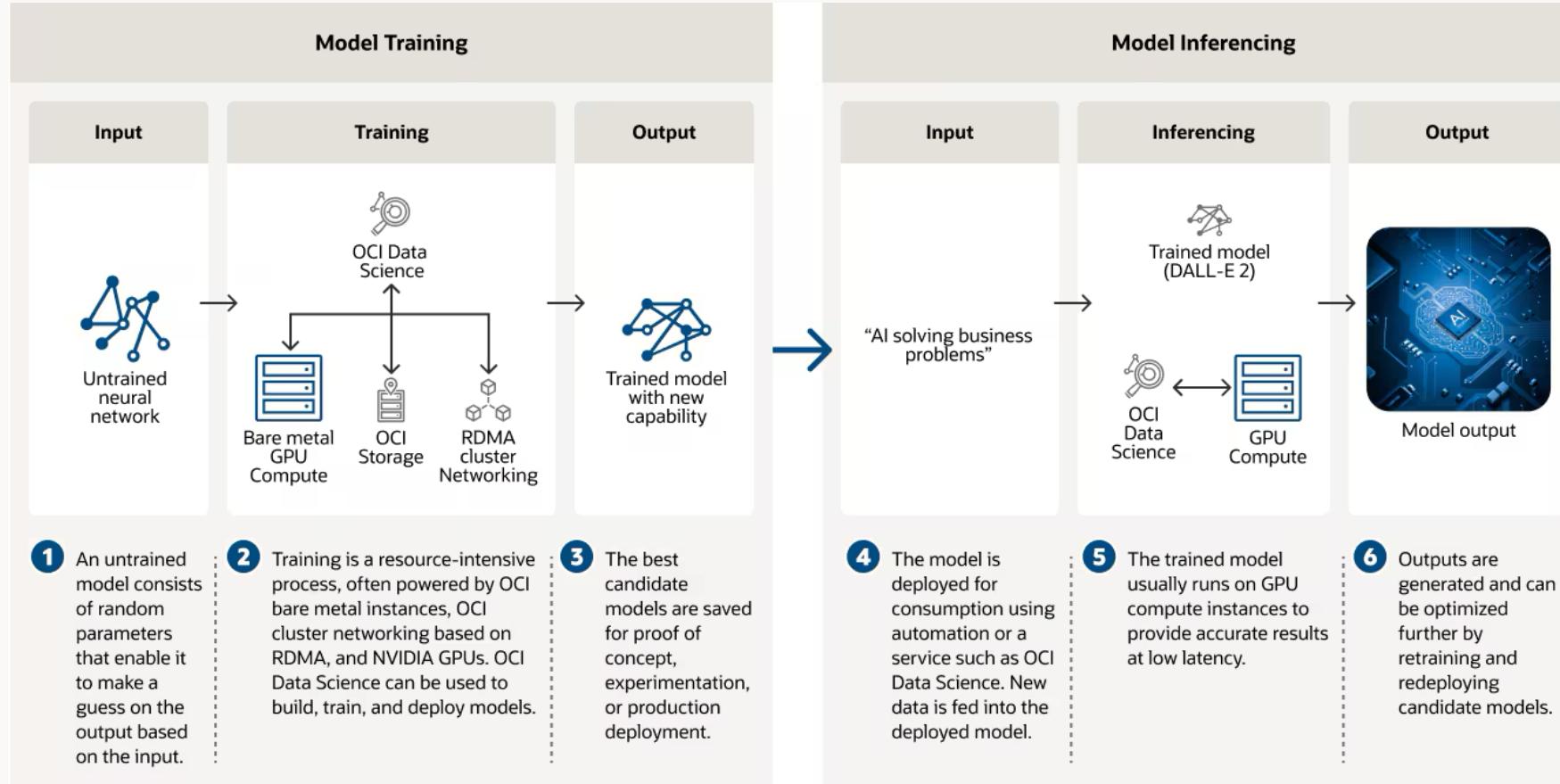


Oracle Cloud Infrastructure from core to edge

Unified experience from cloud to edge

OCI AI Services

OCI AI Service: Bare Metal Cluster for AI/ML



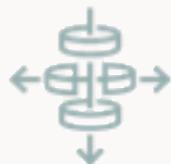
*With the RDMA and GPU superclusters on OCI, customers can implement inside their own **Private AI Cluster**, with their AI workload, either on **Model Training** and/or **Model inferencing** with one unified management experience and talent supports, at any scale.*

OCI AI Service: Prebuilt Services

AI for everyone



Prebuilt for enterprise
needs



Customizable for
your use case



Generative AI

Cognitive AI

OCI AI Service: Prebuilt Models

The screenshot shows the OCI AI Service interface with the following components:

- Left Sidebar (Generative AI):** Includes links for Overview, Playground, Chat (selected), Generation, Summarization, Embedding, Dedicated AI clusters, Custom models, and Endpoints.
- Search Bar:** Search resources, services, documentation, and Marketplace.
- Region Selection:** US Midwest (Chicago).
- Chat Interface:** Chat tab selected. Model dropdown set to xai.grok-3. Example dropdown set to Choose View code button. Type a message... input field. Submit and Clear chat buttons. Request id: ⓘ
- Prebuilt Models Grid:** A grid of four sections: Cohere Models, Meta Models, xAI Models, Embed Models, and Rerank Model.

 - Cohere Models:** Cohere Command A (New), Cohere Command R (08-2024), Cohere Command R+ (08-2024)
 - Meta Models:** Meta Llama 4 Maverick (New), Meta Llama 4 Scout (New), Meta Llama 3.3 (70B), Meta Llama 3.2 90B Vision, Meta Llama 3.2 11B Vision, Meta Llama 3.1 (405B), Meta Llama 3.1 (70B), Meta Llama 3 (70B)
 - xAI Models:** xAI Grok 3, xAI Grok 3 Mini, xAI Grok 3 Fast, xAI Grok 3 Mini Fast
 - Embed Models:** Cohere Embed 4 (New), Cohere Embed English Image 3, Cohere Embed English Light Image 3, Cohere Embed Multilingual Image 3, Cohere Embed Multilingual Light Image 3, Cohere Embed English 3, Cohere Embed English Light 3, Cohere Embed Multilingual 3, Cohere Embed Multilingual Light 3
 - Rerank Model:** Cohere Rerank 3.5 (New)

- Page Bottom:** Terms of Use and Privacy, Cookie Preferences, Copyright © 2025, Oracle and/or its affiliates. All rights reserved., Why don't I see Redwood? | Redwood preview toggle switch.

OCI AI Services: Dedicated AI Cluster for Models Customization

The screenshot shows the OCI AI Services interface. On the left, there's a sidebar with options like Overview, Playground, Chat, Generation, Summarization, Embedding, and Dedicated AI clusters. Under 'Dedicated AI clusters', 'Custom models' is selected. The main content area has a title 'Cluster Network for NVIDIA H100/A100'. It contains two sections: 'Each GPU has its own RDMA connection' with a list of bullet points about H100 and A100 bandwidth, and 'Cluster networking powers OCI super clusters' with a list of bullet points about GPU scaling. At the bottom, there are 'Create' and 'Cancel' buttons.

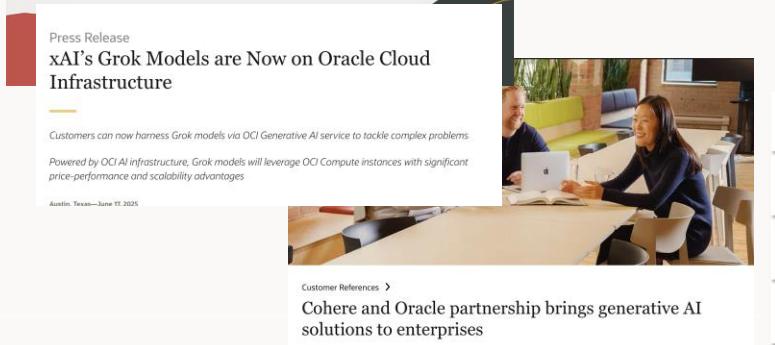
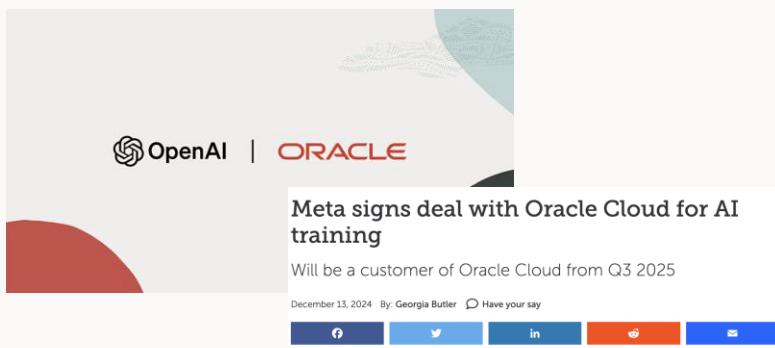
The screenshot shows the 'Create model' wizard. The first step, 'Model definition', is selected. It includes fields for 'Compartment' (set to 'fulli'), 'Name' (optional), 'Version' (optional), and 'Description' (optional). There's also a 'Show advanced options' link. At the bottom, there are 'Next' and 'Cancel' buttons, along with links for 'Terms of Use and Privacy' and 'Cookie Preferences'. A copyright notice at the bottom right states 'Copyright © 2025, Oracle and/or its affiliates. All rights reserved.' and a 'Give us feedback | Redwood preview' button.

With clicks, customers can allocate the state-of-the-art RDMA GPU clusters to train, fine tune their model

Oracle Partnership and Integration

Oracle Partnership and Multi Cloud Strategy

Partnership



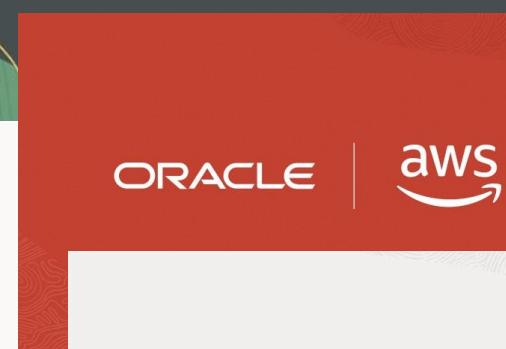
Integration



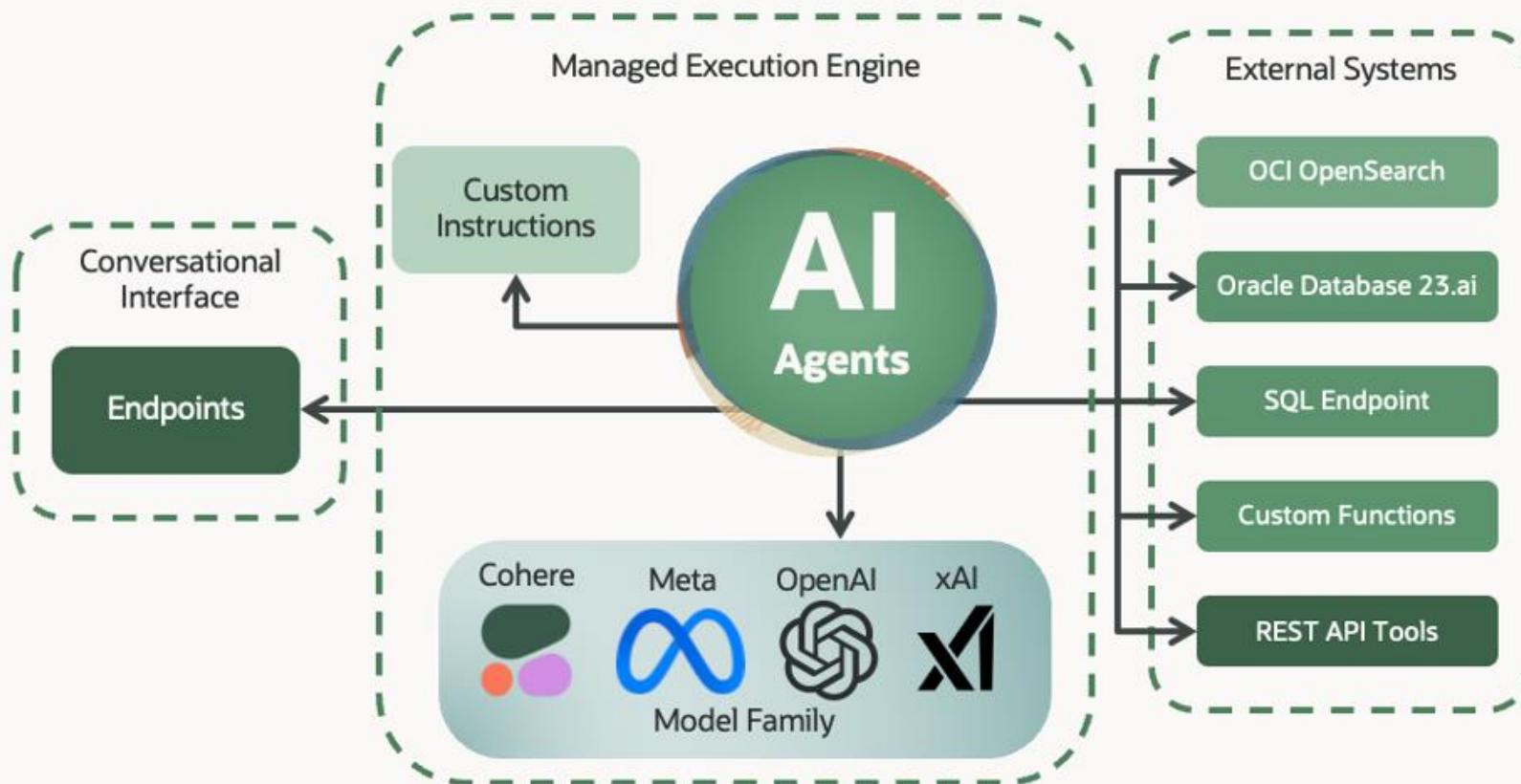
CNCF Project → OCI AI services



Multi Cloud



One More Thing for Agentic AI



For Agentic AI Developers

- Wide model choices
- High performance infra for model training
- AI database integrated
- Embracing cloud native architecture
- Multicloud with low latency between CSPs



OCI AI Infra Demos



Deploy HPC and AI Cluster on OCI: *HPC on Click*

Link: <https://github.com/oracle-quickstart/oci-hpc>

Stack to create an HPC cluster.

Policies to deploy the stack:

```
allow service compute_management to use tag-namespace in tenancy
allow service compute_management to manage compute-management-family in tenancy
allow service compute_management to read app-catalog-listing in tenancy
allow group user to manage all-resources in compartment compartmentName
```

Policies for autoscaling or resizing:

As described when you specify your variables, if you select instance-principal as way of authenticating you make sure your generate a dynamic group and give the following policies to it:

```
Allow dynamic-group instance_principal to read app-catalog-listing in tenancy
Allow dynamic-group instance_principal to use tag-namespace in tenancy
```

- **Open Source**
- Open technology (terraform and ansible)
- Graphical wizard, zero coding needed
- Storage configuration
- Autoscaling with SLURM or PBS Pro
- Monitoring portal
- **RDMA autoconfigured**

Create stack

Welcome!

You're here because you clicked a button to deploy cloud resources, using the [package](#) identified below.

Package URL: <https://github.com/oracle-quickstart/oci-hpc/archive/refs/heads/master.zip>

I have reviewed and accept the [Oracle Terms of Use](#).

Stack information

HPC cluster
Automated HPC cluster deployment

Working directory
oci-hpc-master

Custom providers

Use custom Terraform providers
[Store custom Terraform providers in a bucket](#).

Next Cancel

Deploy Kubernetes with RDMA and GPU on OCI: *GPU RDMA on Click*

Link: <https://github.com/oracle-quickstart/oci-hpc-oke>

The screenshot shows two parallel interfaces. On the left is the GitHub README page for the 'oci-hpc-oke' repository, which contains instructions for deploying a Kubernetes cluster using the Oracle Cloud Resource Manager. A red arrow points from the 'Deploy to Oracle Cloud' button on this page to the 'Create stack' wizard on the right. The right side is the Oracle Cloud Resource Manager's 'Create stack' interface, where a stack for 'OKE: GPU Quickstart (Infrastructure)' is being configured. The wizard steps are: 1. Stack information, 2. Configure variables, 3. Review. The 'Stack information' step is currently active, showing the package URL as <https://github.com/oracle-quickstart/oci-hpc-oke/releases/download/v25.5.1/oke-rdma-quickstart-v25.5.1.zip>. A checkbox indicates acceptance of the Oracle Terms of Use.

Deploy the cluster using the Oracle Cloud Resource Manager

You can easily deploy the cluster using the Deploy to Oracle Cloud button below.

Deploy to Oracle Cloud

For the image ID, use the ID of the image that you imported in the previous step.

The template will deploy a `bastion` instance and an `operator` instance. The `bastion` instance provides access to the OKE cluster. You can connect to the `operator` instance via SSH with the command `ssh ubuntu@<operator IP>`.

You can also find this information under the Application information tab in the Oracle Cloud Infrastructure console.

Wait until you see all nodes in the cluster

```
kubectl get nodes
```

NAME	STATUS	ROLES	AGE	VERSION
10.0.103.73	Ready	<none>	2d23h	v1.25.6
10.0.127.206	Ready	node	2d3h	v1.25.6
10.0.127.32	Ready	node	2d3h	v1.25.6
10.0.83.93	Ready	<none>	2d23h	v1.25.6
10.0.96.82	Ready	node	2d23h	v1.25.6

- **Open Source**
- Open technology (terraform and ansible)
- Graphical wizard, zero coding needed
- Storage configuration
- Monitoring portal
- **RDMA autoconfigured**
- **GPU enabled in OKE**

Deploy AI Blueprint on OCI: *AI Frameworks on Click*

Link: <https://github.com/oracle-quickstart/oci-ai-blueprints>

README UPL-1.0 license Security

OCI AI Blueprints

Deploy, scale, and monitor AI workloads with the OCI AI Blueprints platform, and reduce your GPU onboarding time from weeks to minutes.

OCI AI Blueprints is a streamlined, no-code solution for deploying and managing Generative AI workloads on Kubernetes Engine (OKE). By providing opinionated hardware recommendations, pre-packaged software stacks, and out-of-the-box observability tooling, OCI AI Blueprints helps you get your AI applications running quickly and efficiently—without wrestling with the complexities of infrastructure decisions, software compatibility, and MLOps best practices.

[Install OCI AI Blueprints](#)

Table of Contents

Getting Started

- [Install AI Blueprints](#)
- [Access AI Blueprints Portal and API](#)

About OCI AI Blueprints

- [What is OCI AI Blueprints?](#)
- [Why use OCI AI Blueprints?](#)
- [Features](#)
- [List of Blueprints](#)
- [FAQ](#)
- [Support & Contact](#)

API Reference

- [API Reference Documentation](#)

Cloud Search resources, services, documentation, and Marketplace US Midwest (Chicago)

Create stack

Welcome!

You're here because you clicked a button to deploy cloud resources, using the package identified below.

Package URL: https://github.com/oracle-quickstart/oci-ai-blueprints/releases/download/v1.0.4/v1.0.4_app.zip

I have reviewed and accept the [Oracle Terms of Use](#).

Stack information [i](#)

OCI AI Blueprints

This OKE Starter Kit is a comprehensive collection of Terraform scripts designed to automate the creation and deployment of resources to easily run AI/ML workloads in OCI. This stack provisions resources such as an OKE cluster, an ATP database instance, and other essential components, which enable you to deploy AI/ML workloads with a simple UI or an API. The kit includes services such as Grafana and Prometheus for infrastructure monitoring, MLFlow for tracking experiment-level metrics, and KEDA for dynamic auto-scaling based on AI/ML workload metrics rather than infrastructure metrics.

Working directory `oci_ai_blueprints_terraform`

The file path to the directory from which to run Terraform

Custom providers

Use custom Terraform providers
[Store custom Terraform providers in a bucket](#)

Name *Optional* `v1.0.4_app.zip-20250708152942`

Description *Optional*

ORACLE