ORACLE

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!*

ORACLE

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

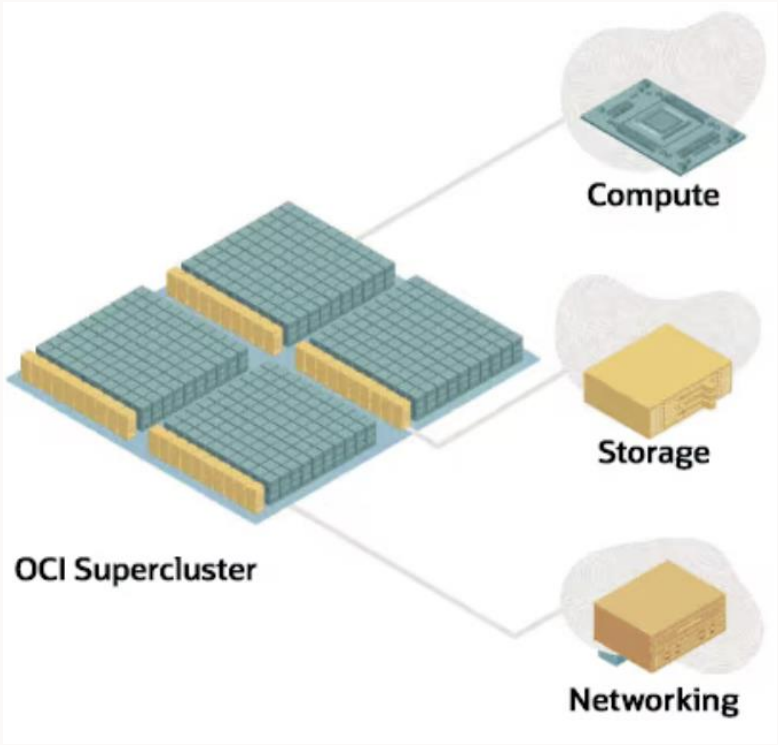
- On-premises-like performance and stability
  - Bare metal servers: no hypervisor, no virtualization penalty
  - ***Real HPC RDMA network***: up to 3,200Gbps/server bandwidth + low µs latency
  - ***NVMe on every node***
- With the flexibility, cost savings, and functionality of the cloud
- Partnership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

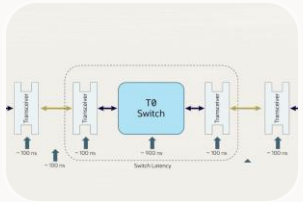
Massive capacity and high-throughput storage

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

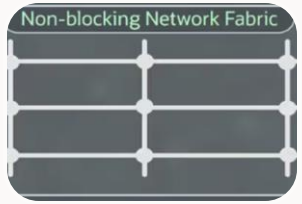
Ultrafast nonblocking network

- 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

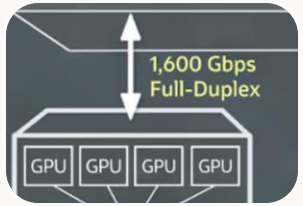
Features



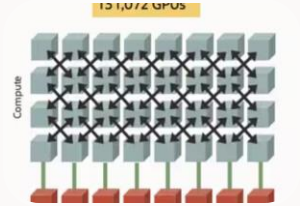
2.5 us latency



nonblocking RDMA

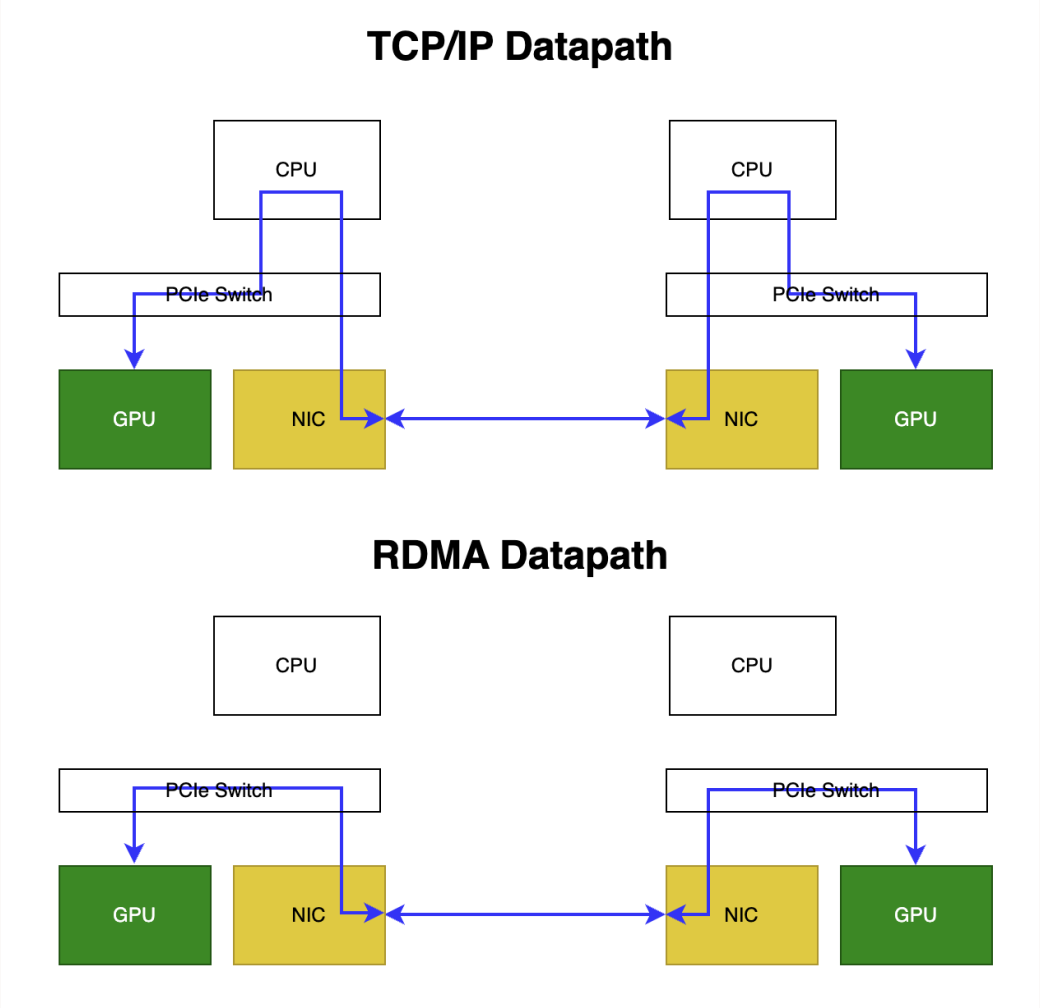


3.2Tb/s Bandwidth



131,072 GPUs Scalability

Why RDMA? Bypass CPU, Low Latency and Full Bandwidth

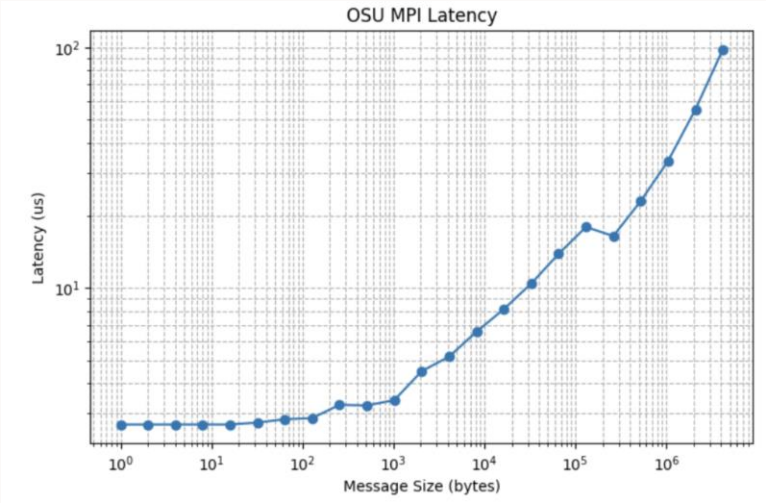


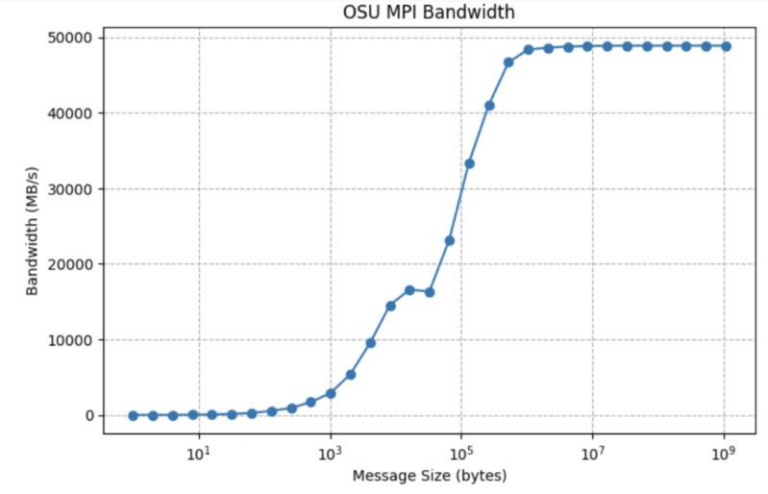
RDMA vs TCP/IP

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.

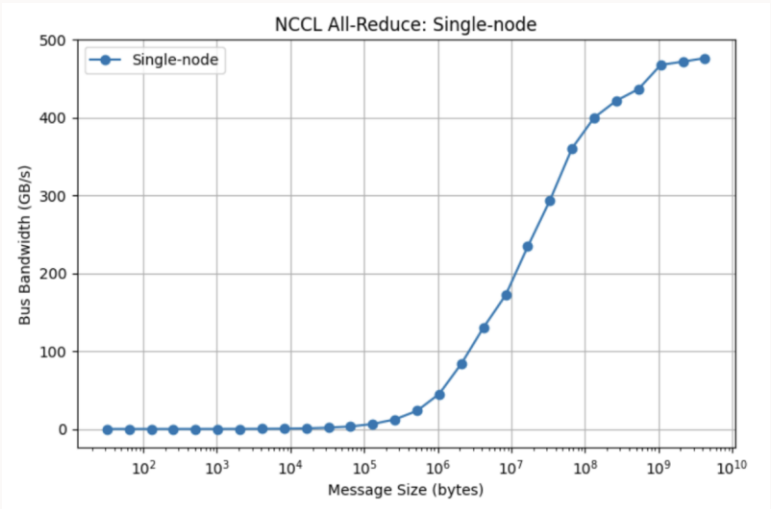
Metrics Talk: the State-of-the-Art Performance of OCI Supercluster

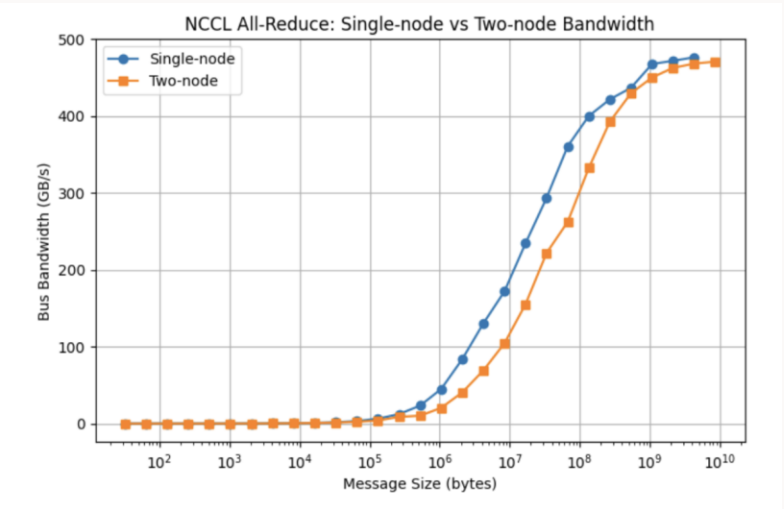




OSU MPI Benchmarking

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

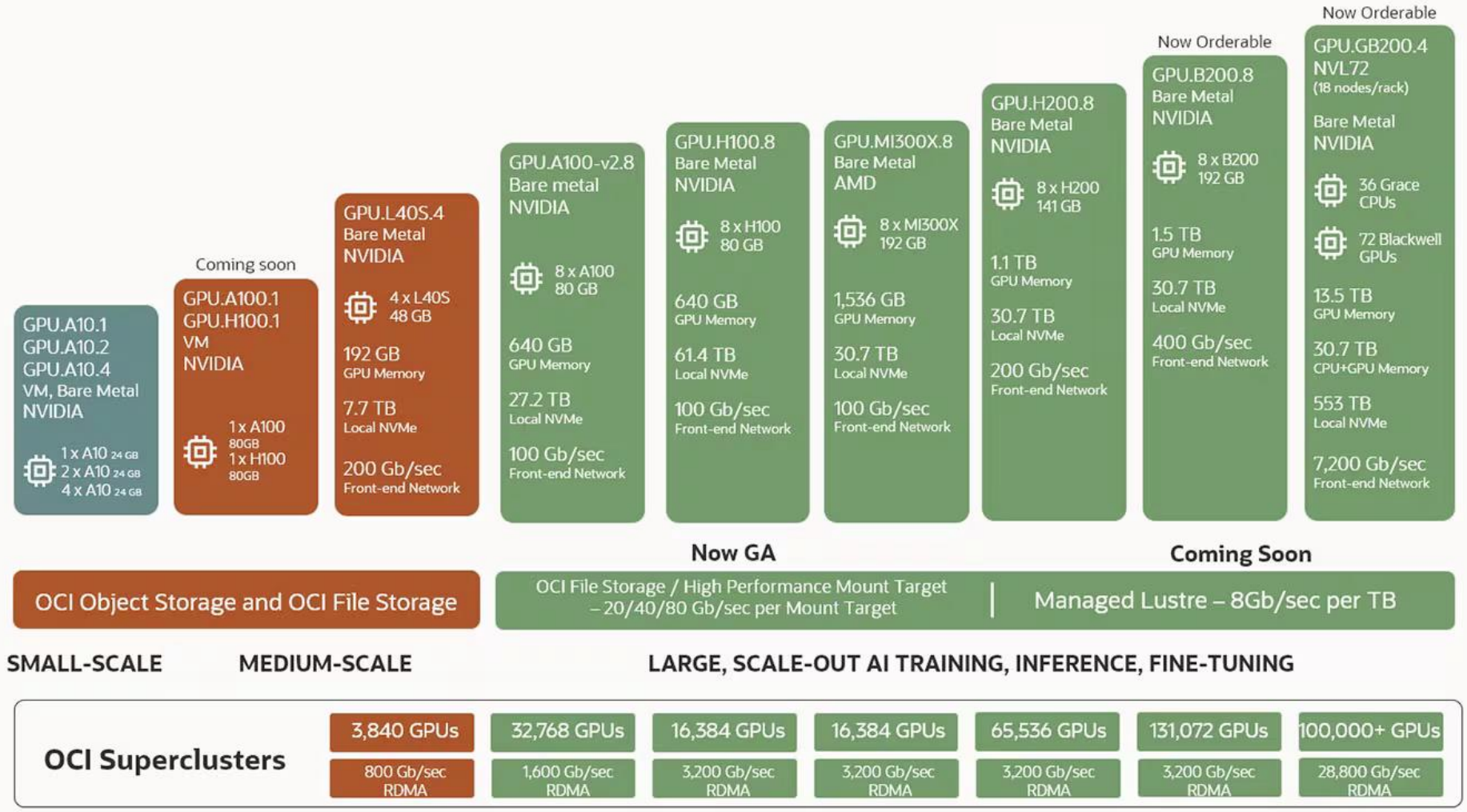




NCCL MPI Benchmarking

- 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

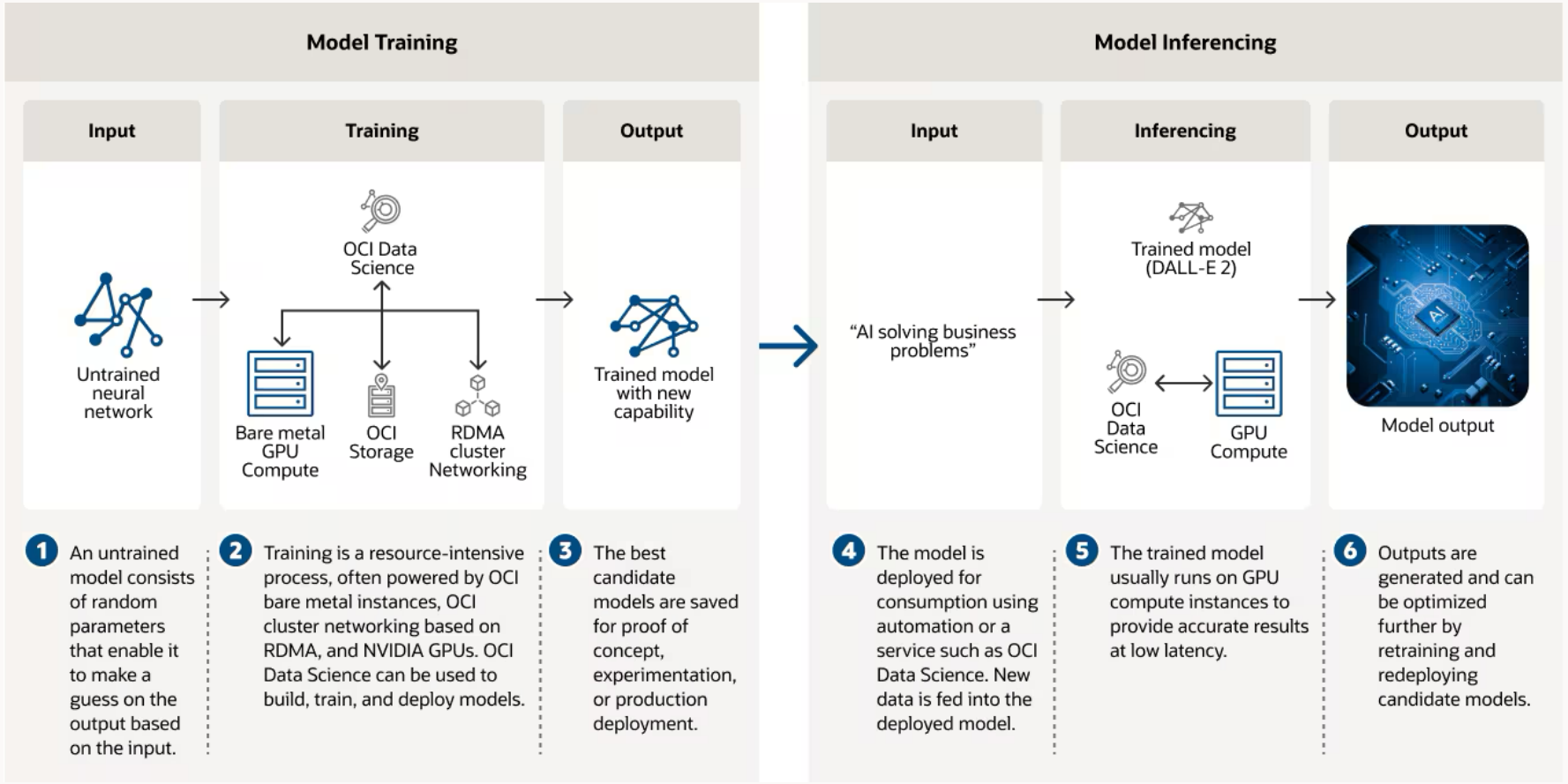
Oracle Roving Edge Infrastructure	Oracle Compute Cloud@Customer	OCI Dedicated Region	OCI Public Region
			
8 – 48 OCPUs	552 – 6600+ OCPUs	Full Spectrum of OCI Region IaaS	
NVIDIA L4 GPUs	NVIDIA L40S GPUs	NVIDIA H100, A100, L40S, A10, V100, AMD MI300X	
OCI IaaS, OKE, select OCI Services, Oracle & 3rd party marketplace solutions		150 + OCI Services	
OCI APIs, CLI, SDKs, Tools, and User Experience			
Unified control plane across all deployment options – Develop Once, Deploy Anywhere			
Globally consistent cloud economics model			

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

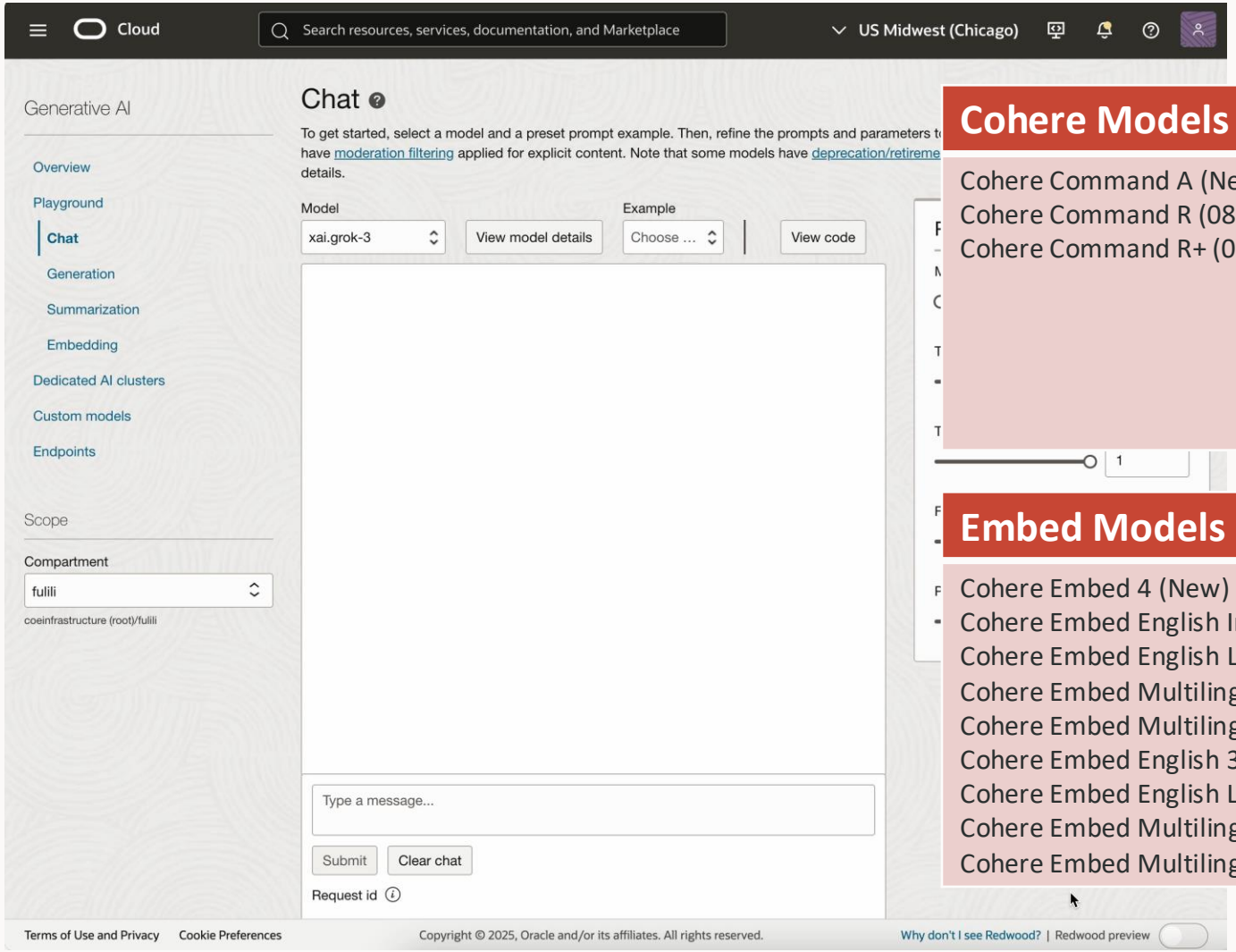
Prebuilt for enterprise needs

Customizable for your use case

Generative AI

Cognitive AI

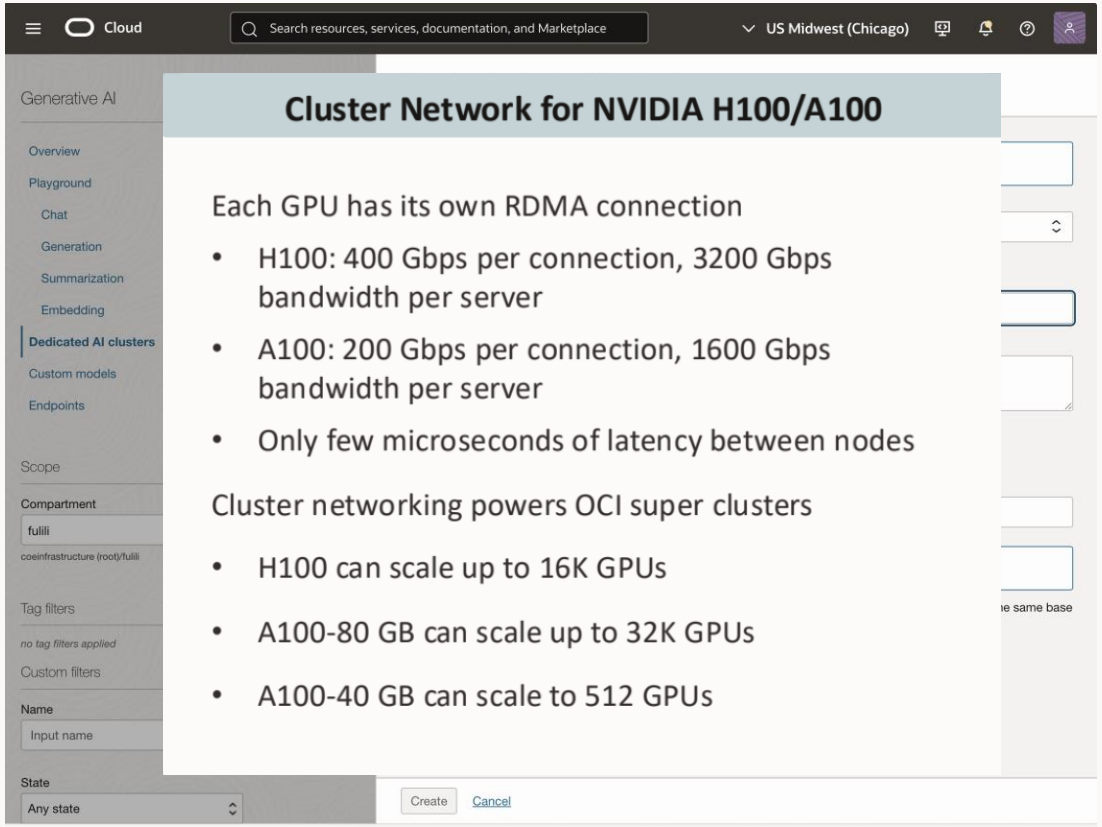
OCI AI Service: Prebuilt Models

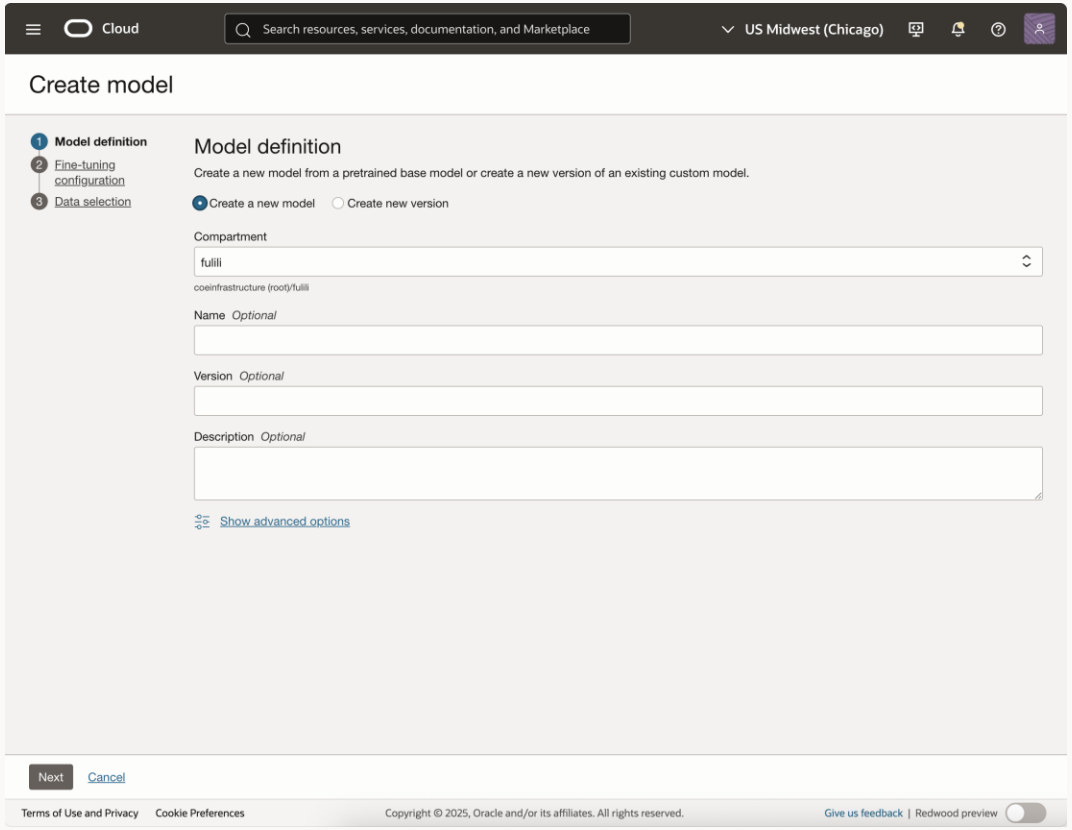


Cohere Models	Meta Models	xAI Models
Cohere Command A (New) Cohere Command R (08-2024) Cohere Command R+ (08-2024)	Meta Llama 4 Maverick (New) Meta Llama 4 Scout (New) Meta Llama 3.3 (70B) Meta Llama 3.2 90B Vision Meta Llama 3.2 11B Vision Meta Llama 3.1 (405B) Meta Llama 3.1 (70B) Meta Llama 3 (70B)	xAI Grok 3 xAI Grok 3 Mini xAI Grok 3 Fast xAI Grok 3 Mini Fast

Embed Models	Rerank Model
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	Cohere Rerank 3.5 (New)

OCI AI Services: Dedicated AI Cluster for Models Customization





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

OpenAI | ORACLE

Meta signs deal with Oracle Cloud for AI training

Will be a customer of Oracle Cloud from Q3 2025

December 13, 2024 By: Georgia Butler Have your say

...o use the company's artificial intelligence (AI)

...a company will use Oracle Cloud for the training an...

...a customer in the company's recent earnings call.

ORACLE + xAI

Press Release

xAI's Grok Models are Now on Oracle Cloud Infrastructure

Customers can now harness Grok models via OCI Generative AI service to tackle complex problems

Powered by OCI AI infrastructure, Grok models will leverage OCI Compute instances with significant price-performance and scalability advantages

Austin, Texas—June 17, 2025

Customer References >

Cohere and Oracle partnership brings generative AI solutions to enterprises

Integration

Oracle CONNECT

Oracle Database 23ai sets new industry standard for a modern AI database

CNCF Project → **OCI AI services**

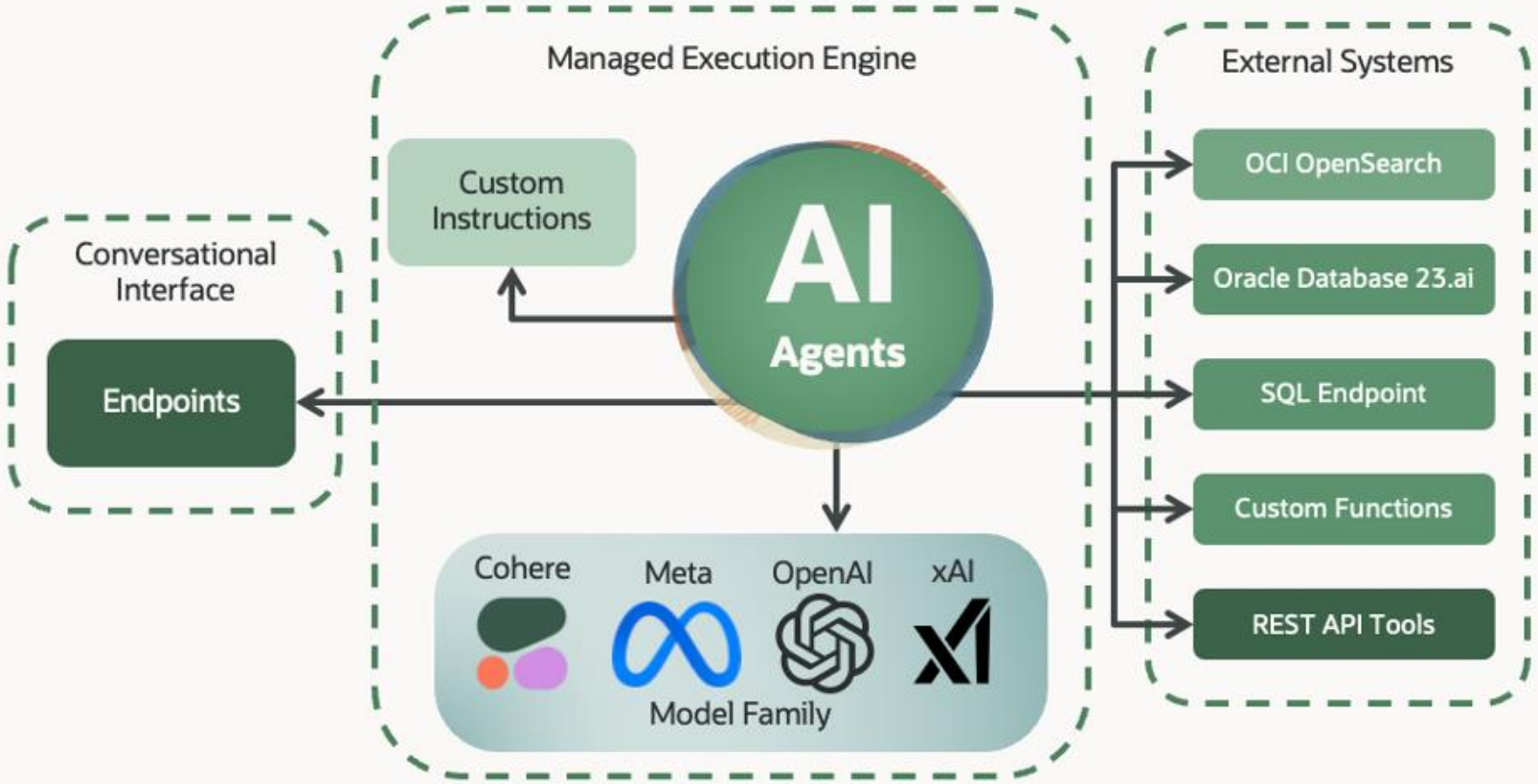
Multi Cloud

ORACLE | Microsoft

ORACLE | aws

ORACLE | Google Cloud

One More Thing for Agentic AI



For Agentic AI Developers

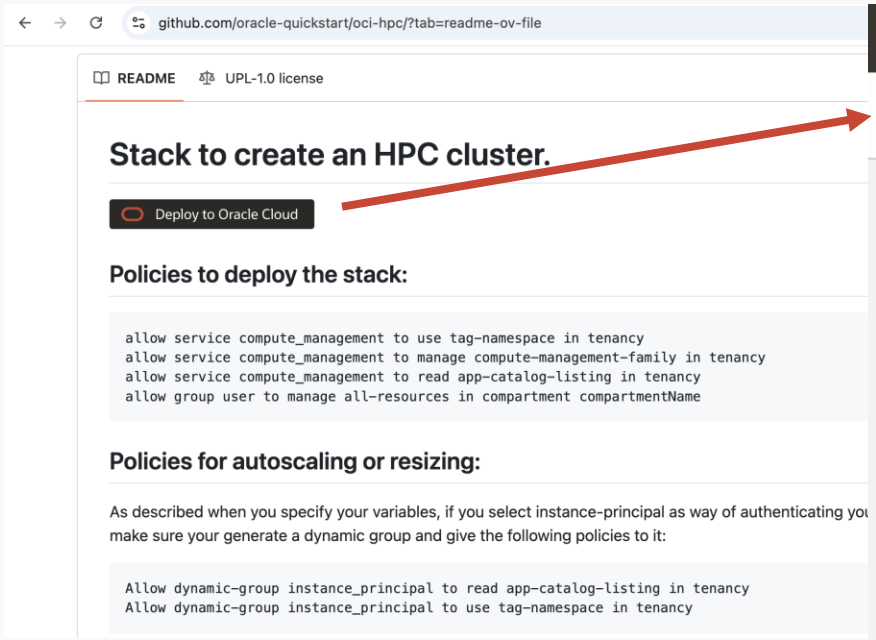
- Wide model choices
- High performance infra for model training
- AI database integrated
- Embracing cloud native architecture
- Multicloud with low latency between CSPs

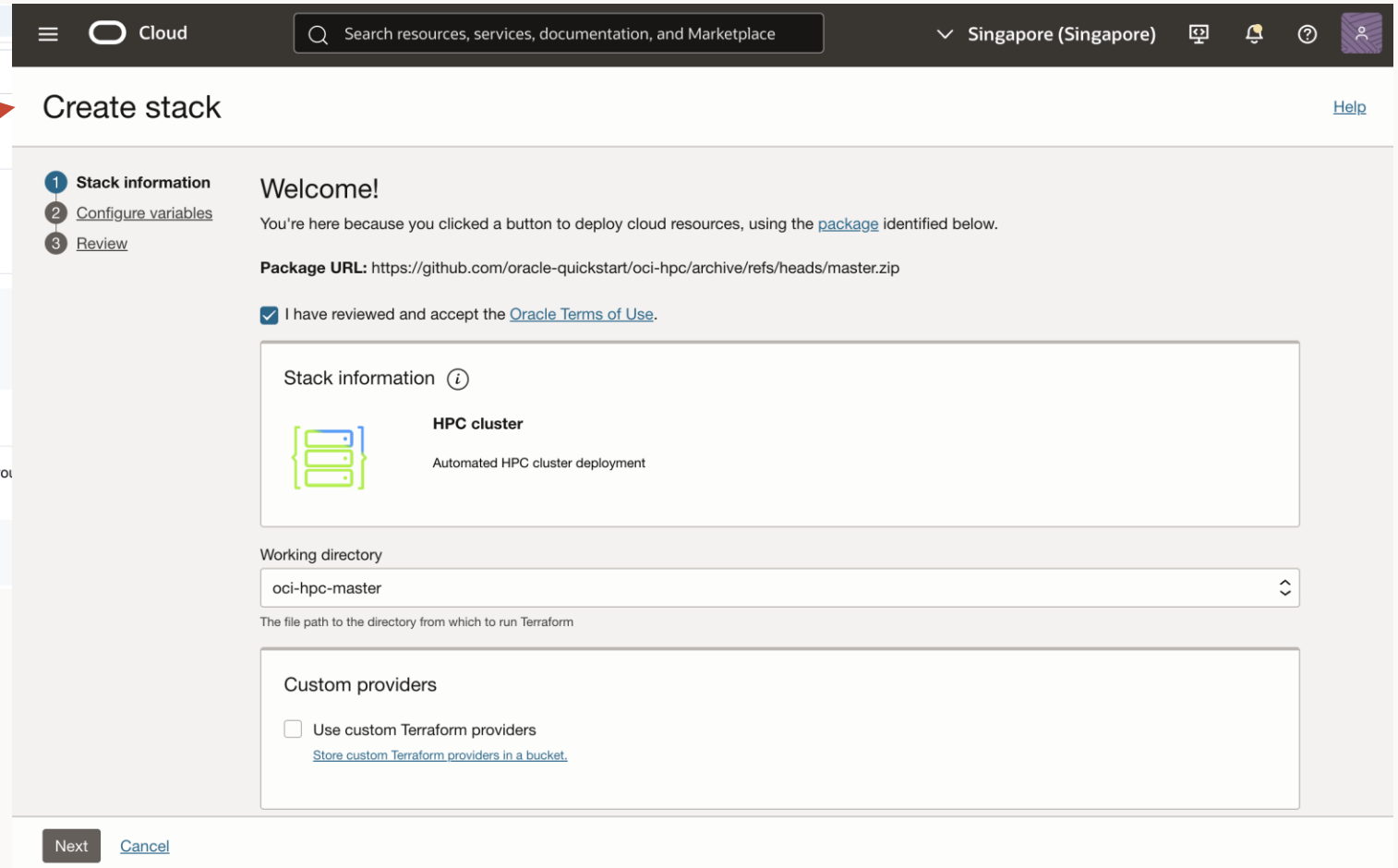
ORACLE

OCI AI Infra Demos

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

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

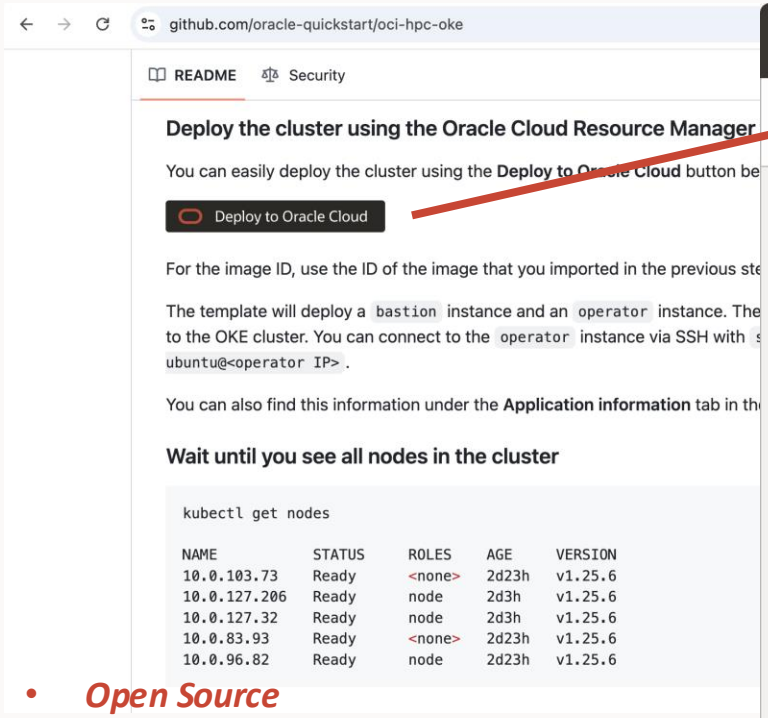


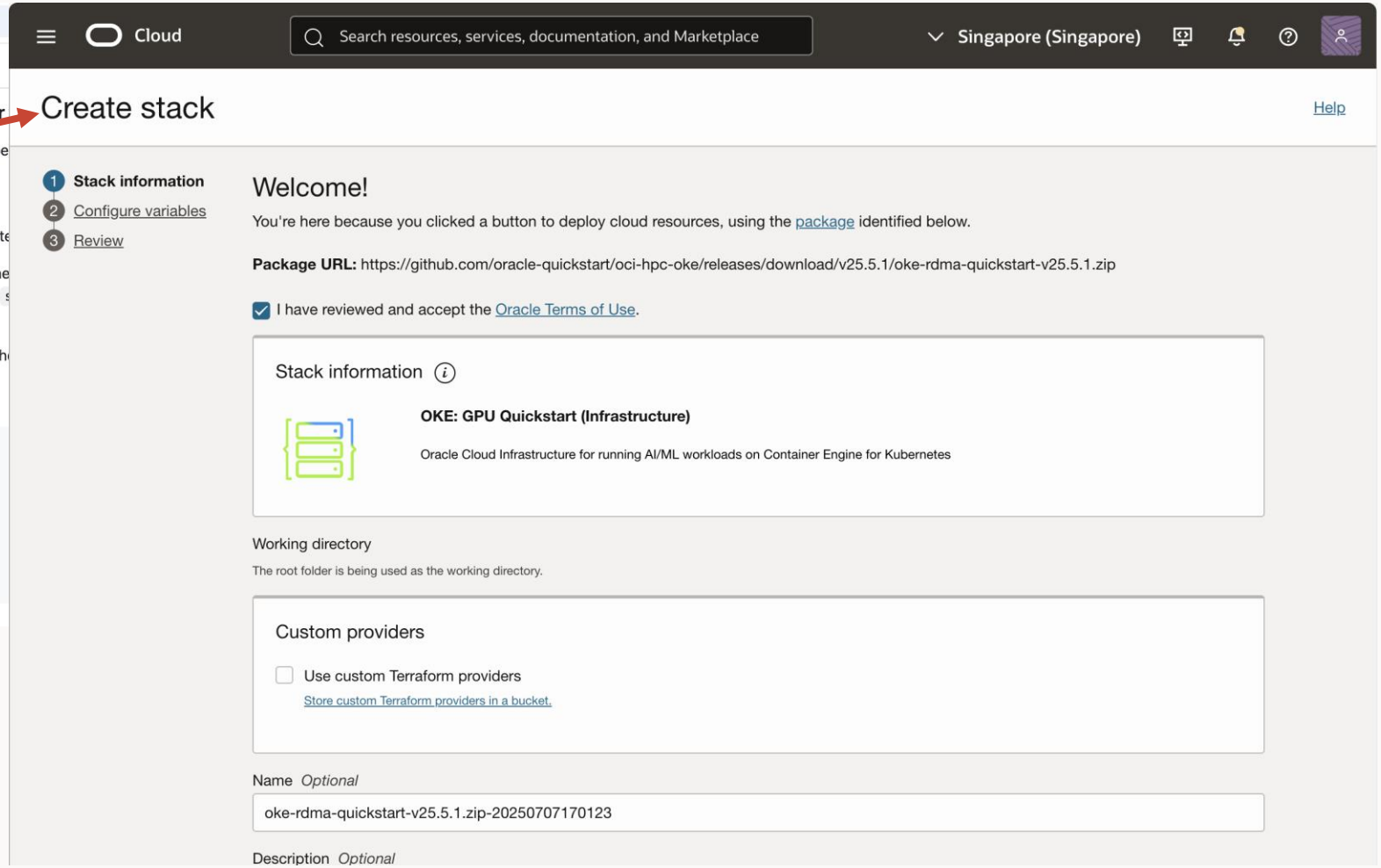


- ***Open Source***
- Open technology (terraform and ansible)
- Graphical wizard, zero coding needed
- Storage configuration
- Autoscaling with SLURM or PBS Pro
- Monitoring portal
- ***RDMA autoconfigured***

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

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





- ***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