ORACLE

Accelerating AI Workloads with OCI Supercluster

Oracle's AI infrastructure is scalable, performant, and deployable anywhere

A history of AI infrastructure innovation

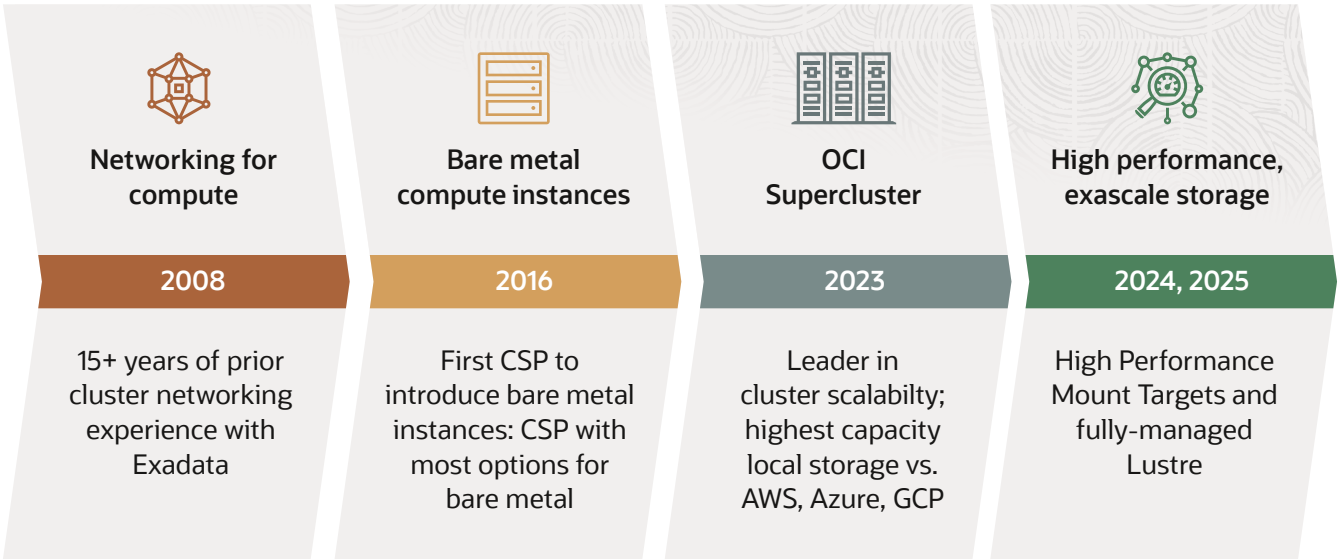
Enabling scalable and cost-efficient AI

Generative AI demands cost-efficient, high performance infrastructure to scale and accelerate model training and deployment.

Oracle has long been a leader in high performance infrastructure, which is now essential for AI workloads. When it launched in 2008, Oracle Exadata featured unique and revolutionary Remote Direct Memory Access (RDMA). In 2016, Oracle Cloud Infrastructure (OCI) was the first to offer bare metal compute instances to give customers maximum performance and control.

OCI Supercluster, launched in 2023, delivers one of the highest performance and lowest-cost GPU clusters in the world. Hardware-enabled RDMA on nonblocking networks, substantial local non-volatile memory express (NVMe) storage, and bare metal compute provide the ideal environment for AI training.

In 2024, Oracle launched a petabyte-scale, managed high performance mount target file system, useful for training large-scale large language models (LLMs). In addition, we recently announced managed Lustre file system service for trillion-parameter models that require the highest performance storage.



Innovating with OCI Supercluster

OCI Supercluster provides today's most advanced and highest performance cloud environment for building and deploying AI, scaling up to an industry-leading 131,072 NVIDIA Blackwell GPUs.* At maximum scale, it offers the performance of the previous generation, as well as 25X less energy consumption and 25X better TCO.

OCI Superclusters with NVIDIA GB200 (Grace Blackwell) and B200 (Blackwell) are generally available for AI training and inference. They provide up to 30X improvement in AI inference performance and 4X in AI training performance compared to previous-generation OCI Supercluster with NVIDIA H100 Tensor Core GPUs.

In addition, OCI Supercluster with NVIDIA H200 Tensor Core GPUs features instances with 76% more high-bandwidth GPU memory capacity and 40% more GPU memory bandwidth than the previous-generation NVIDIA H100 GPU instances, improving LLM inference performance by up to 1.9X. With double the front-end network throughput for data ingestion and retrieval at 200 Gb/sec per instance, data transfer to and from the cluster is also dramatically improved to further accelerate AI model training.

Companies innovating with OCI

CSM

Common Sense Machines leveraged hands-on expertise from OCI to support GenAI startups.
Learn more

Twelve Labs

Twelve Labs gained 5X to 10X training efficiencies, helping it go to market sooner than expected.
Learn more

SUNO

Suno AI composes quality music and audio with AI foundation models trained on high performance and scalable OCI Supercluster.
Learn more

Scalability for CSP 1: 20,000 NVIDIA H200 GPUs; scalability for CSP 2 and CSP 3: not publicly available

AI infrastructure for any workload

OCI supports AI workloads of all sizes. This chart illustrates OCI Supercluster's scalability and the range of instances that are available or coming soon. OCI can be deployed anywhere, with the broadest set of options in the industry, including OCI Dedicated Region, Oracle Alloy, and more.

How Oracle AI infrastructure stands out

Superior scalability

With a maximum supercluster scalability of 131,072 NVIDIA GPUs available in 2025, OCI can meet the needs of the largest generative AI deployments in the world.

Bare metal instances

In addition to a purpose-built, dedicated network based on ethernet and NVIDIA Quantum-2 InfiniBand, Oracle AI infrastructure offers bare metal compute instances that remove the overhead of a virtualization layer for better performance.

Leading distributed cloud

OCI is a leading provider of distributed cloud infrastructure in the public cloud, sovereign and government clouds, on-premises data centers, and partner data centers.

Tailored support

OCI includes 24/7 operations support and dedicated cloud engineers with AI expertise to guide you in deploying, troubleshooting, and managing AI infrastructure.

Pricing

With uniform pricing for services globally, including those running on AI infrastructure, GPU instances on OCI can cost significantly less than those from other CSPs.

Why Uber chose Oracle

Uber

Uber uses OCI for more than a million trips per hour

Uber has modernized its application tier and AI infrastructure and migrated much of its operational big data and streaming stack to OCI to help it drive profitable growth, deliver new products to market, and accelerate innovation.

14 Million

Predictions per second

1 Million

Trips every hour

Why Zoom chose Oracle

zoom

Zoom AI Companion revolutionizes the way organizations work

Zoom, the AI-first work platform for human connection, scaled its users while improving performance and saving on infrastructure costs. Zoom AI Companion, the company's personal AI assistant, can help users find information, draft emails and chat messages, summarize meetings and make them more actionable, improve brainstorming, kickstart content generation, and more, right in the Zoom Workplace app.

> "By harnessing OCI's AI inference capabilities, Zoom is able to deliver accurate results at low latency, empowering users to collaborate seamlessly, communicate effortlessly, and boost productivity, efficiency, and potential like never before.

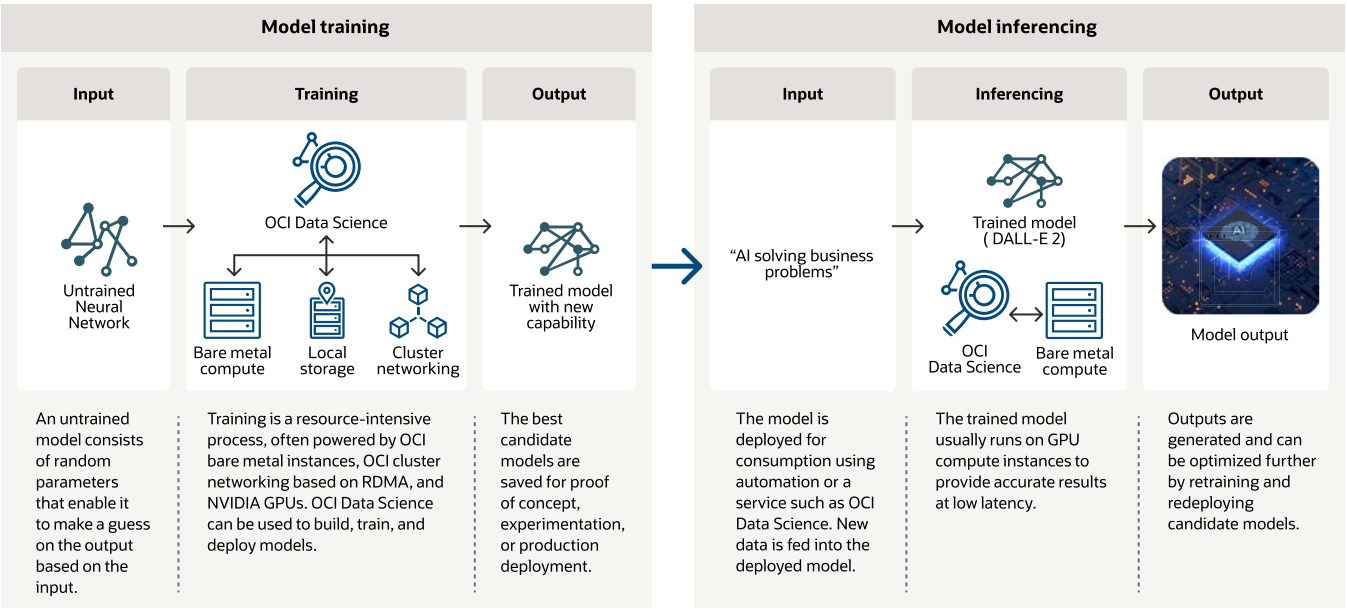
Bo Yan
Head of AI, Zoom

Learn more about Zoom's AI-First Culture on OCI

AI infrastructure in action: Explore top use cases

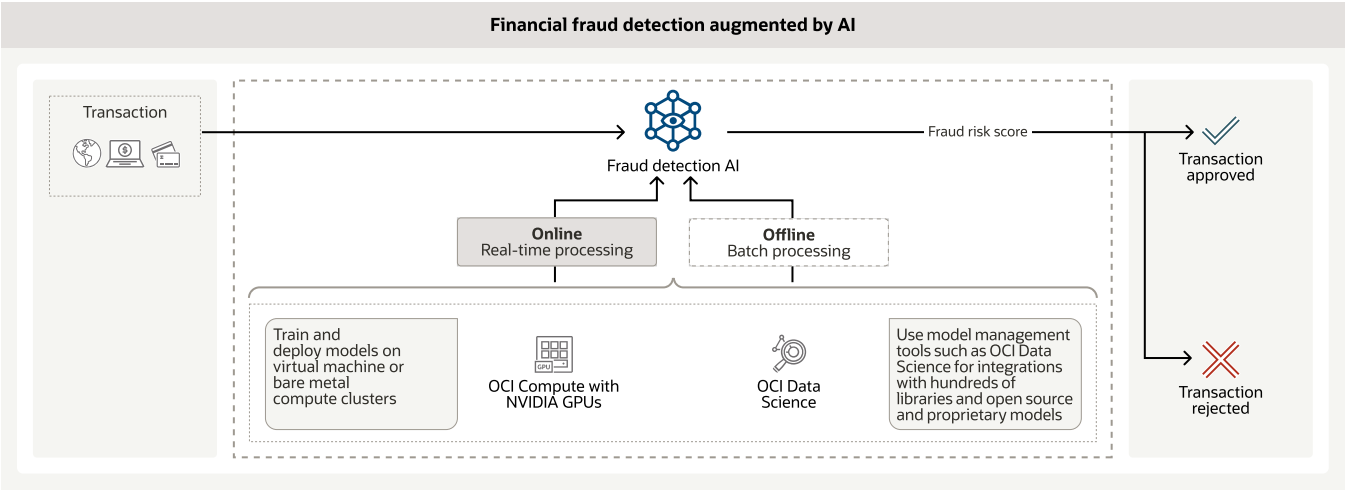
LLM training and inferencing

AI startups and enterprises can train AI models using bare metal GPU instances and ultrafast cluster networking. Leverage developer services for Oracle-managed and self-managed Kubernetes orchestration with support for GPUs, APIs for prebuilt AI services, and third-party tools such as PyTorch, TensorFlow, JAX, Kueue, and more.



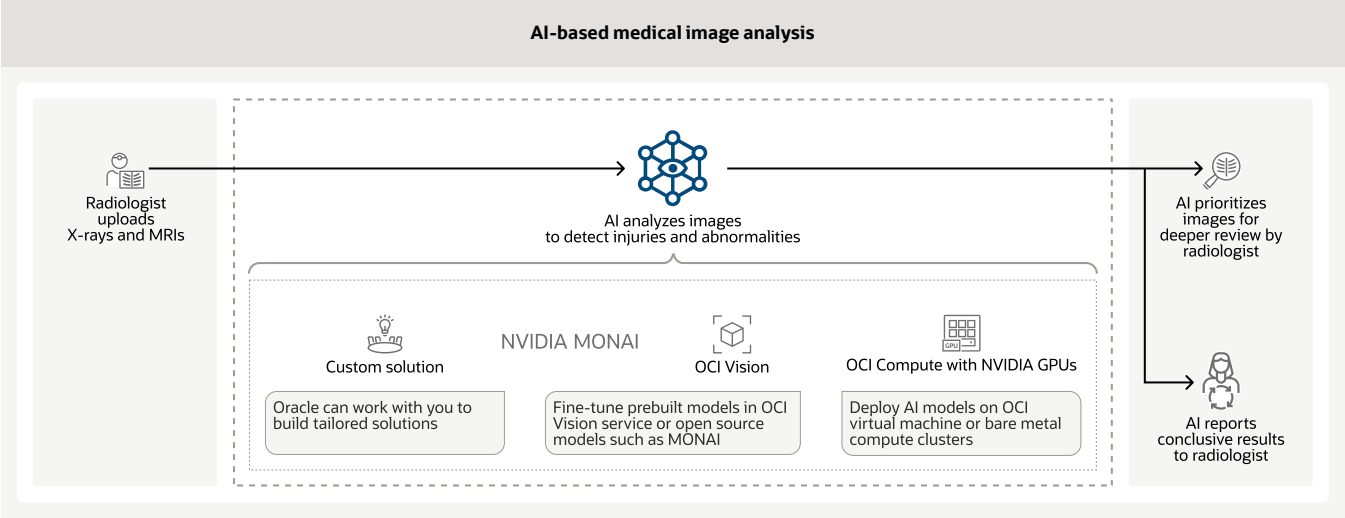
Fraud detection augmented by AI

Protecting billions of financial transactions daily requires enhanced AI tools that have analyzed extensive historical transaction data. AI models running on OCI and accelerated by NVIDIA and AMD GPUs help financial institutions mitigate fraud.



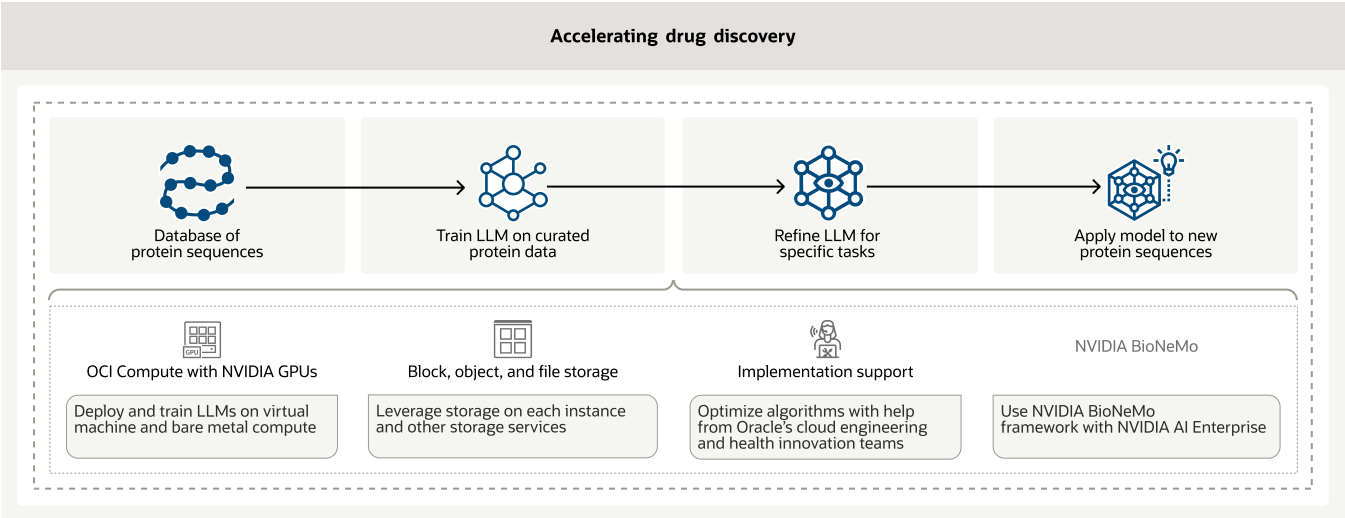
AI-based medical image analysis

Trained models analyze X-rays, CT scans, and MRIs to help prioritize images that need immediate review by a radiologist and report conclusive results from others.



Drug discovery accelerated by AI

By leveraging AI infrastructure and analytics, researchers can accelerate drug discovery, a process that previously took years. Additionally, AI workflow management tools such as NVIDIA BioNeMo™ can help researchers curate and preprocess data.



Distributed cloud and sovereign AI

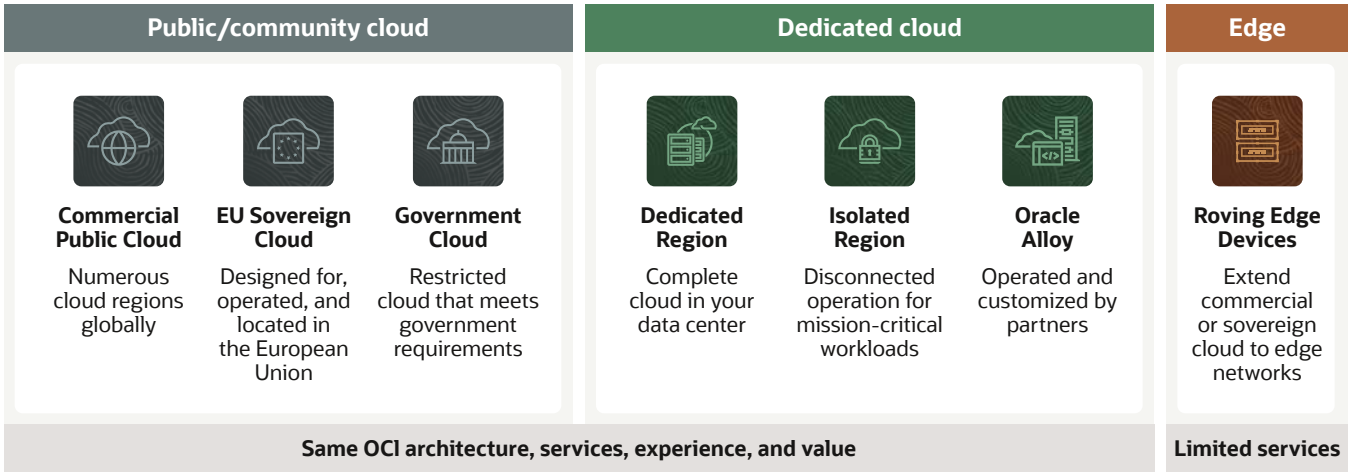
OCI was named a Leader in the Gartner Magic Quadrant for Distributed Hybrid Infrastructure and rated highly for both the Sovereignty and Distributed Cloud Architecture critical capabilities.

With OCI, you can run workloads in more locations than with any other hyperscaler; there are 160+ live or planned regions. Further, OCI offers unique multicloud capabilities, including Oracle Database@Azure, Oracle Database@Google Cloud, and Oracle Database@AWS, that let you bring your most useful data to the hyperscale cloud of your choice.

OCI Dedicated Region and Oracle Alloy support sovereign AI, which refers to governmental or organizational control over AI technologies and associated data. With Oracle, you decide how and where AI technologies are deployed and operated, including specifying the hardware and software infrastructure as well as the policies and personnel used to operate the AI technologies and protect your data.

You can achieve AI sovereignty with Oracle AI and OCI distributed cloud solutions, which give you the power to determine where AI workloads run and how data and systems are managed.

Deploy AI applications anywhere with OCI distributed cloud



Partnering with NVIDIA for success

NVIDIA software combined with OCI's AI cloud infrastructure provides a broad, easily accessible portfolio of tools and services for AI training and inferencing at scale. The following NVIDIA software is available with OCI:

- NVIDIA NIM™ microservices for accelerating generative AI model deployment anywhere
- NVIDIA DGX™ Cloud, a high performance, fully managed AI platform that provides optimized accelerated computing clusters
- NVIDIA AI Enterprise, an end-to-end software platform for production AI
- NVIDIA RAPIDS™, open source libraries and APIs that use GPUs to accelerate data science and model training, including Apache Spark workloads
- NVIDIA TensorRT-LLM, a library for optimizing LLM inference–it provides state-of-the-art optimizations, including custom attention kernels, in-flight batching, paged KV caching, quantization (FP8, INT4 AWQ, INT8 SmoothQuant, ++) and much more, to perform inference efficiently on NVIDIA GPUs
- NVIDIA Triton Inference Server, open source software that standardizes AI model deployment and execution across every workload
- NVIDIA BioNeMo™, a collection of programming tools, libraries, and models for computational drug discovery

See how you can accelerate AI with Oracle Cloud and NVIDIA.

> "The limitless opportunities for AI-driven innovation are helping transform virtually every business. NVIDIA's collaboration with Oracle Cloud Infrastructure puts the extraordinary supercomputing performance of NVIDIA's accelerated computing platform within reach of every enterprise.
>
> **Justin Boitano**
> Vice President of Enterprise AI, NVIDIA

Now it's your turn to build the future

See what customers have achieved with OCI
Learn how companies are building and running applications on Oracle's scalable, secure, highly available, fault-tolerant, and high performance cloud environment.
Read their stories

See how much you can save with OCI
Oracle Cloud pricing is simple and consistently low worldwide, supporting a wide range of use cases. To estimate your low rate, check out the cost estimator and configure the services to suit your needs.
Estimate your costs

Get started with AI infrastructure
Learn more about RDMA cluster networking, GPU instances, bare metal servers, and more.

Explore AI infrastructure

Connect with us

Call +1.800.ORACLE1 or visit oracle.com

Outside North America, find your local office at oracle.com/contact