Oracle and NVIDIA Collaborate to Help Enterprises Accelerate Agentic AI Inference

Oracle Database and NVIDIA AI Integrations Make It Easier for Enterprises to Quickly and Easily Harness Agentic AI

GTC—Oracle and NVIDIA today announced a first-of-its-kind integration between NVIDIA accelerated computing and inference software with Oracle's AI infrastructure, and generative AI services, to help organizations globally speed creation of agentic AI applications.

The new integration between Oracle Cloud Infrastructure (OCI) and the NVIDIA AI Enterprise software platform will make 160+ AI tools and 100+ NVIDIA NIM™ microservices natively available through the OCI Console. In addition, Oracle and NVIDIA are collaborating on the no-code deployment of both Oracle and NVIDIA AI Blueprints and on accelerating AI vector search in Oracle Database 23ai with the NVIDIA cuVS library.

"Oracle has become the platform of choice for both AI training and inferencing, and this partnership enhances our ability to help customers achieve greater innovation and business results," said Safra Catz, CEO of Oracle. "NVIDIA's offerings, paired with OCI's flexibility, scalability, performance and security, will speed AI adoption and help customers get more value from their data."

"Oracle and NVIDIA are perfect partners for the age of reasoning — an AI and accelerated computing company working with a key player in processing much of the world's enterprise data," said Jensen Huang, founder and CEO of NVIDIA. "Together, we help enterprises innovate with agentic AI to deliver amazing things for their customers and partners."

Purpose-Built Solutions to Meet Enterprise AI Needs
Reducing the time it takes to deploy reasoning models, NVIDIA AI Enterprise will be natively available through the OCI Console, enabling customers to quickly and easily access AI tools including NVIDIA NIM — a set of 100+ optimized, cloud-native inference microservices for leading AI models, including the latest NVIDIA Llama Nemotron models for advanced AI reasoning.

NVIDIA AI Enterprise will be available as a deployment image for OCI bare-metal instances and Kubernetes clusters using OCI Kubernetes Engine. OCI Console customers benefit from direct billing and customer support through Oracle.

Organizations can deploy OCI's 150+ AI and cloud services with NVIDIA accelerated computing and NVIDIA AI Enterprise in the data center, the public cloud or at the edge. This offering provides an integrated AI stack to help address data privacy, sovereign AI and low-latency requirements.

Biotechnology company Soley Therapeutics is deploying OCI AI Infrastructure, NVIDIA AI Enterprise and NVIDIA Blackwell GPUs to build its AI drug discovery platform to unlock possible treatments for complex diseases by capturing, decoding and interpreting cellular language to forecast cell fate.

"We believe in the potential of AI in developing new solutions that can help deliver treatments for cancer and other complex diseases," said Yerem Yeghiazarians, cofounder and CEO of Soley Therapeutics. "The combination of OCI and NVIDIA delivers a full-stack AI solution, providing us the storage, compute, software tools and support necessary to innovate faster with petabytes of data in developing our AI drug discovery platform."

AI Deployment at Scale With Tailored Blueprints
OCI AI Blueprints provide no-code deployment recipes that enable customers to quickly run AI workloads without having to make decisions about the software stack or manually provision the infrastructure. The blueprints offer clear hardware recommendations for NVIDIA GPUs, NIM microservices and prepackaged observability tools, helping enterprises accelerate their AI projects from weeks to minutes.

In addition, NVIDIA Blueprints provide developers with a unified experience across the NVIDIA stack, providing reference workflows for enterprise AI use cases. Using NVIDIA Blueprints, organizations can build and operationalize custom AI applications with NVIDIA AI Enterprise and NVIDIA Omniverse™ software, application programming interfaces and microservices. For example, developers can begin with an NVIDIA AI Blueprint for a customer service AI assistant and customize it for their own use.

To simplify the development, deployment and scale-out of advanced physical AI and simulation applications and workflows, the NVIDIA Omniverse platform and NVIDIA Isaac Sim™ development workstations and Omniverse Kit App Streaming are expected to be available on Oracle Cloud Infrastructure Marketplace later this year, preconfigured with compute bare-metal instances accelerated by NVIDIA L40S GPUs.

Pipefy, an AI-powered automation platform for business process management, uses an inference blueprint for document preprocessing and image processing.

“We embraced OCI AI Blueprints to spin up NVIDIA GPU nodes and deploy multimodal large language models quickly for document- and image-processing use cases,” said Gabriel Custodio, principal software engineer at Pipefy. “Using these prepackaged and verified blueprints, deploying our AI models on OCI is now fully automated and significantly faster.”

Real-Time AI Inference With NVIDIA NIM in OCI Data Science
To further accelerate enterprise AI adoption and help enable quick AI deployments with minimal setup, data scientists can access pre-optimized NVIDIA NIM microservices directly in OCI Data Science. This supports real-time AI inference use cases without the complexity of managing infrastructure.

To help maintain data security and compliance, the models run in the customer’s OCI tenancy. Customers can purchase the models through a flexible, pay-as-you-go, hourly pricing model or apply their Oracle Universal Credits.

Organizations can use this integration to deploy inference endpoints with preconfigured, optimized NIM inference engines in minutes, rapidly accelerating time to value for use cases such as AI-powered assistants, real-time recommendation engines and copilots. In addition, this allows customers to start using the integration for smaller workloads and seamlessly scale to enterprise-wide deployments.

NVIDIA Accelerated Computing Platform Turbocharges AI Vector Search in Oracle Database 23ai
Oracle and NVIDIA are working together to accelerate the creation of vector embeddings and vector indexes — compute-intensive portions of AI Vector Search workloads in Oracle Database 23ai — using NVIDIA GPUs and NVIDIA cuVS.

Organizations can enable vector embedding through bulk vectorization of large volumes of input data such as text, images and videos, as well as the fast creation and maintenance of vector indexes. With NVIDIA-accelerated AI Vector Search, Oracle Database customers can significantly improve the performance of their AI pipelines to help support high-volume AI vector workloads.

DeweyVision provides advanced computer vision and artificial intelligence capabilities to turn media into data, making it accessible, searchable, discoverable, retrievable and actionable. DeweyVision uses Oracle Database 23ai on Oracle Autonomous Database for its AI-powered, no-code warehousing tools. These tools enable production professionals to connect their workflows and edit video footage quickly by cataloging footage in minutes and providing intuitive search capabilities.

“Oracle Database 23ai with AI Vector Search can significantly increase Dewey’s search performance while increasing the scalability of the DeweyVision platform,” said Majid Bemanian, CEO of DeweyVision. “Using NVIDIA GPUs to create the vector embeddings that we load into Oracle Database accelerates our platform’s ingestion of new data, while Autonomous Database and the converged capabilities of Oracle Database 23ai will help reduce our operational costs as we grow and open new opportunities. We believe that the combination of DeweyVision, Oracle Database 23ai and NVIDIA GPUs running in OCI will help us achieve our goal of becoming Hollywood’s data warehouse.”

NVIDIA Blackwell on OCI Enables AI Anywhere
Oracle and NVIDIA continue to evolve AI infrastructure with new NVIDIA GPU types across OCI’s public regions, government clouds, sovereign clouds, OCI Dedicated Region, Oracle Alloy, OCI Compute Cloud@Customer and OCI Roving Edge Devices.

This includes NVIDIA Quantum-2 InfiniBand cluster network environments, NVIDIA Spectrum™ Ethernet switches and optimized NVIDIA NVLink™ and NVLink Switch functionality for some of the largest AI superclusters in the market.

OCI will offer NVIDIA GB200 NVL72 systems on OCI Supercluster — generally available soon with up to 131,072 NVIDIA GPUs — and is taking orders for one of the largest AI supercomputers in the cloud with NVIDIA Blackwell Ultra GPUs.

OCI will be among the first cloud service providers to offer the next generation of the NVIDIA Blackwell accelerated computing platform. Built on the groundbreaking Blackwell architecture introduced a year ago, Blackwell Ultra includes the NVIDIA GB300 NVL72 rack-scale solution and the NVIDIA HGX™ B300 NVL16 system. The GB300 NVL72 delivers 1.5x more AI performance than the NVIDIA GB200 NVL72, as well as increases Blackwell’s revenue opportunity by 50x for AI factories, compared with those built with NVIDIA Hopper™.

SoundHound, a global leader in conversational intelligence, offers voice and conversational AI solutions, powering voice-related experiences in millions of products from global brands. Its voice AI platform runs on OCI, processing billions of queries annually, and uses NVIDIA GPUs to provide customers with fast and accurate voice services.

“SoundHound has developed a long-term relationship with OCI, and we believe our ongoing collaboration will play a key role in supporting future growth,” said James Hom, chief product officer of SoundHound AI. “NVIDIA GPUs will greatly accelerate training for our next generation of voice AI.”

Additional Resources

- Learn more about OCI AI infrastructure.
- Learn more about Oracle Cloud Infrastructure.
- Learn more about Oracle expanding its distributed cloud capabilities with NVIDIA AI Enterprise.
- Learn more about OCI AI infrastructure for OCI Dedicated Region and Oracle Alloy.
- Read the OCI Supercluster technical blog and watch the video.

About Oracle Distributed Cloud
Oracle's distributed cloud delivers the benefits of cloud with greater control and flexibility. Oracle's distributed cloud lineup includes:

- Public cloud: Hyperscale public cloud regions serve any size of organization, including those requiring strict EU sovereignty controls. See the full list of regions.
- Dedicated cloud: Customers can run all OCI cloud services in their own data centers with OCI Dedicated Region, while partners can resell OCI cloud services and customize the experience using Oracle Alloy. Oracle also operates separate U.S., U.K. and Australian government clouds, as well as isolated cloud regions for national security purposes. Each of these products provides a full cloud and AI stack that customers can deploy as a sovereign cloud.
- Hybrid cloud: OCI delivers key cloud services on-premises via Oracle Exadata Cloud@Customer and Compute Cloud@Customer and is already managing deployments in over 60 countries. Additionally, OCI Roving Edge Infrastructure, which consists of multiple configurations of ruggedized and portable high-performance devices, helps customers leverage remote AI inferencing at the edge.
- Multicloud: OCI is physically deployed within all the hyperscale cloud providers, including AWS, Google Cloud and Microsoft Azure, providing low-latency, natively integrated Oracle database services, including Oracle Database@AWS, Oracle Database@Azure, Oracle Database@Google Cloud and Oracle HeatWave on AWS. Oracle Interconnect for Microsoft Azure and Oracle Interconnect for Google Cloud allows customers to combine key capabilities from across clouds.

About NVIDIA
NVIDIA (NASDAQ: NVDA) is the world leader in accelerated computing.

About Oracle
Oracle offers integrated suites of applications plus secure, autonomous infrastructure in the Oracle Cloud. For more information about Oracle (NYSE: ORCL), please visit us at oracle.com.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, availability, and performance of NVIDIA's products, services, and technologies; and the collaboration between NVIDIA and Oracle and the benefits and impact thereof are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

Oracle, Java, MySQL and NetSuite are registered trademarks of Oracle Corporation. NetSuite was the first cloud company—ushering in the new era of cloud computing.

© 2025 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA HGX, NVIDIA Hopper, NVIDIA Isaac Sim, NVIDIA NIM, NVIDIA Omniverse, NVIDIA Spectrum-X and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and/or other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability, and specifications are subject to change without

notice.

Cliff Edwards
NVIDIA Corporation
+1-415-699-2755
cliffe@nvidia.com
Ben Wolfson
Oracle
+1.510.480.5230
ben.wolfson@oracle.com