Oracle and NVIDIA to Deliver Sovereign AI Worldwide

Oracle and NVIDIA collaborate to deliver accelerated computing and generative AI services that establish digital sovereignty and manage proprietary national and personal data

Oracle adopts NVIDIA Grace Blackwell across OCI Supercluster, OCI Compute, and NVIDIA DGX Cloud on OCI

AUSTIN, Texas and SAN JOSE, Calif.—GTC—March 21, 2024 — Oracle and NVIDIA today announced an expanded collaboration to deliver sovereign AI solutions to customers around the world. Oracle's distributed cloud, AI infrastructure, and generative AI services, combined with NVIDIA's accelerated computing and generative AI software, are enabling governments and enterprises to deploy AI factories.

These AI factories can run cloud services locally, and within a country's or organization's secure premises with a range of operational controls, supporting sovereign goals of diversifying and boosting economic growth.

"As AI reshapes business, industry, and policy around the world, countries and organizations need to strengthen their digital sovereignty in order to protect their most valuable data," said Safra Catz, CEO of Oracle. "Our continued collaboration with NVIDIA and our unique ability to deploy cloud regions quickly and locally will ensure societies can take advantage of AI without compromising their security."

"In an era where innovation will be driven by generative AI, data sovereignty is a cultural and economic imperative," said Jensen Huang, founder and CEO of NVIDIA. "Oracle's integrated cloud applications and infrastructure, combined with NVIDIA accelerated computing and generative AI services, create the flexibility and security nations and regions require to control their own destiny."

Turnkey Solutions to Help Customers Meet Data Sovereignty

The combination of NVIDIA's full-stack AI platform with Oracle's Enterprise AI – deployable across OCI Dedicated Region, Oracle Alloy, Oracle EU Sovereign Cloud, and Oracle Government Cloud – offers customers a state-of-the-art AI solution that provides greater control over operations, location, and security to help support digital sovereignty.

Countries across the globe are increasingly investing in AI infrastructure that can support their cultural and economic ambitions. Across 66 cloud regions in 26 countries, customers can access more than 100 cloud and AI services spanning infrastructure and applications to support IT migration, modernization, and innovation.

The companies' combined offerings can be deployed via the public cloud or in a customer's data center in specific locations, with flexible operational controls. Oracle is the only hyperscaler capable of delivering AI and full cloud services locally, anywhere. OCI services and pricing are consistent across deployment types to simplify planning, portability, and management.

Oracle's cloud services leverage a range of NVIDIA's stack, including NVIDIA accelerated computing infrastructure and the NVIDIA AI Enterprise software platform, including newly ~~just~~ announced NVIDIA NIM™ inference microservices, which are built on the foundation of NVIDIA inference software such as NVIDIA TensorRT™, NVIDIA TensorRT-LLM, and NVIDIA Triton Inference Server™.

Sovereign AI Pioneers

Avaloq, a leader in wealth management technology, selected OCI Dedicated Region to bring a complete OCI cloud region into its own data center.

"OCI Dedicated Region aligns with our commitment to ensure maximum control over data residency while providing access to the latest cloud infrastructure," said Martin Büchi, chief technology officer at Avaloq. "This supports us as we continue to drive the digital transformation of banks and wealth managers."

TEAM IM, a leading New Zealand information management services provider, chose Oracle Alloy to build New Zealand's first locally owned and operated hyperscale cloud known as TEAM Cloud.

"Organizations in New Zealand are increasingly eager to harness the power of the cloud while safeguarding the integrity of their data within their own shores by leveraging a unique hyperscale cloud solution," said Ian Rogers, chief executive officer of TEAM IM. "With Oracle Alloy and the possibility of integrating the NVIDIA AI platform into our cloud services, we've been able to become a cloud services provider that can assist public sector, commercial and iwi organizations in navigating the intricacies of the digital landscape and optimizing their digital transformations."

e& UAE, telecom arm of e& group, is collaborating with Oracle to enhance its AI capabilities and intends to deploy NVIDIA H100 Tensor Core GPU clusters within its OCI Dedicated Region.

"OCI will enable us to deploy NVIDIA H100 GPU clusters within our own OCI Dedicated Region, hosted at e& UAE data centers," said Khalid Murshed, chief technology and information officer (CTIO) of e& UAE. "This type of localization will allow us to accelerate AI innovation across the UAE and helps us develop new Gen AI applications and use cases at scale. This is in line with e& UAE's transformation efforts to pioneer innovation and shape the future of technology with our focus on driving excellence in AI to provide unparalleled customer experiences."

OCI Supercluster and OCI Compute Boosted with NVIDIA Grace Blackwell

To help customers address the ever-increasing needs of AI models, Oracle plans to take advantage of the latest NVIDIA Grace Blackwell computing platform, announced today at GTC, across OCI Supercluster and OCI Compute. OCI Supercluster will become significantly faster with new OCI Compute bare metal instances, ultra-low-latency RDMA networking, and high-performance storage. OCI Compute will adopt both the NVIDIA GB200 Grace Blackwell Superchip and the NVIDIA Blackwell B200 Tensor Core GPU.

The NVIDIA GB200 Grace™ Blackwell Superchip will power a new era of computing. GB200 delivers up to 30X faster real-time large language model (LLM) inference, 25X lower TCO, and requires 25X less energy compared to the previous generation of GPUs, supercharging AI training, data processing, and engineering design and simulation. NVIDIA Blackwell B200 Tensor Core GPUs are designed for the most demanding AI, data analytics, and high-performance computing (HPC) workloads.

NVIDIA NIM and CUDA-X™ microservices, including NVIDIA NeMo Retriever for retrieval-augmented generation (RAG) inference deployments, will also help OCI customers bring more insight and accuracy to their generative AI copilots and other productivity tools using their own data.

NVIDIA Grace Blackwell Comes to DGX Cloud on OCI

To meet escalating customer demand for increasingly complex AI models, the companies are adding NVIDIA Grace Blackwell to NVIDIA DGX™ Cloud on OCI. Customers will be able to access new GB200 NVL72 based instances through this co-engineered supercomputing service designed for energy-efficient training and inference in an era of trillion-parameter LLMs.

The full DGX Cloud cluster buildout will include more than 20,000 GB200 accelerators and NVIDIA CX8 InfiniBand networking, providing a highly scalable and performant cloud infrastructure. The cluster will consist of 72 Blackwell GPUs NVL72 and 36 Grace CPUs with fifth-generation NVLink™.

Availability

Oracle and NVIDIA's sovereign AI solutions are available immediately. To learn more, go to the Oracle sovereign AI page.

Additional Resources

- Learn more about OCI Supercluster
- Read more about AI innovators running on OCI
- Learn more about Oracle's distributed cloud strategy

About NVIDIA

Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing infrastructure company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

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 www.oracle.com.

# # #

Media Contact Info

Oracle PR

Carolin Bachmann
carolin.bachmann@oracle.com
+1.415.622.8466

NVIDIA PR

Cliff Edwards
cliffe@nvidia.com
+1.415.699.2755

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features, and availability of NVIDIA's products and technologies, including NVIDIA accelerated computing infrastructure and the NVIDIA AI Enterprise software platform including NVIDIA NIM inference microservices, TensorRT, NVIDIA TensorRT-LLM, NVIDIA

Triton Inference Server; NVIDIA H100 Tensor Core GPU, NVIDIA GB200 Grace Blackwell Superchip, NVIDIA Blackwell B200 Tensor Core GPU, NVIDIA CUDA-X microservices, including NVIDIA NeMo Retriever for retrieval-augmented generation (RAG) inference deployments, NVIDIA DGX Cloud, GB200 NVL72, and NVLink; the benefits and impact of NVIDIA's collaboration with Oracle, and the features and availability of its services and offerings; and innovation being driven by generative AI 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; NVIDIA's reliance on third parties to manufacture, assemble, package and test NVIDIA's products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA's existing product and technologies; market acceptance of NVIDIA's products or NVIDIA partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA's 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.

Trademarks

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.