Oracle Cloud Infrastructure Offers New NVIDIA GPU-Accelerated Compute Instances

NVIDIA H100 Tensor Core GPUs now generally available, and NVIDIA L40S GPUs coming soon on Oracle Cloud Infrastructure.

Author: Dave Salvator

With generative AI and large language models (LLMs) driving groundbreaking innovations, the computational demands for training and inference are skyrocketing.

These modern-day generative AI applications demand full-stack accelerated compute, starting with state-of-the-art infrastructure that can handle massive workloads with speed and accuracy. To help meet this need, Oracle Cloud Infrastructure today announced general availability of NVIDIA H100 Tensor Core GPUs on OCI Compute , with NVIDIA L40S GPUs coming soon.

The OCI Compute bare-metal instances with NVIDIA H100 GPUs, powered by the NVIDIA Hopper architecture , enable an order-of-magnitude leap for large-scale AI and high-performance computing, with unprecedented performance, scalability and versatility for every workload.

Organizations using NVIDIA H100 GPUs obtain up to a 30x increase in AI inference performance and a 4x boost in AI training compared with tapping the NVIDIA A100 Tensor Core GPU . The H100 GPU is designed for resource-intensive computing tasks, including training LLMs and inference while running them.

The BM.GPU.H100.8 OCI Compute shape includes eight NVIDIA H100 GPUs, each with 80GB of HBM2 GPU memory. Between the eight GPUs, 3.2TB/s of bisectional bandwidth enables each GPU to communicate directly with all seven other GPUs via NVIDIA NVSwitch and NVLink 4.0 technology . The shape includes 16 local NVMe drives with a capacity of 3.84TB each and also includes 4th Gen Intel Xeon CPU processors with 112 cores, as well as 2TB of system memory.

In a nutshell, this shape is optimized for organizations' most challenging workloads.

Depending on timelines and sizes of workloads, OCI Supercluster allows organizations to scale their NVIDIA H100 GPU usage from a single node to up to tens of thousands of H100 GPUs over a high-performance, ultra-low-latency network.

The NVIDIA L40S GPU, based on the NVIDIA Ada Lovelace architecture , is a universal GPU for the data center, delivering breakthrough multi-workload acceleration for LLM inference and training, visual computing and video applications. The OCI Compute bare-metal instances with NVIDIA L40S GPUs will be available for early access later this year, with general availability coming early in 2024.

These instances will offer an alternative to the NVIDIA H100 and A100 GPU instances for tackling smaller- to medium-sized AI workloads, as well as for graphics and video compute tasks. The NVIDIA L40S GPU achieves up to a 20% performance boost for generative AI workloads and as much as a 70% improvement in fine-tuning AI models compared with the NVIDIA A100.

The BM.GPU.L40S.4 OCI Compute shape includes four NVIDIA L40S GPUs, along with the latest-generation Intel Xeon CPU with up to 112 cores, 1TB of system memory, 15.36TB of low-latency NVMe local storage for caching data and 400GB/s of cluster network bandwidth. This instance was created to tackle a wide range of use cases, ranging from LLM training, fine-tuning and inference to NVIDIA Omniverse workloads and industrial digitalization, 3D graphics and rendering, video transcoding and FP32 HPC.

This collaboration between OCI and NVIDIA will enable organizations of all sizes to join the generative AI revolution by providing them with state-of-the-art NVIDIA H100 and L40S GPU-accelerated infrastructure.

Access to NVIDIA GPU-accelerated instances may not be enough, however. Unlocking the maximum potential of NVIDIA GPUs on OCI Compute means having an optimal software layer. NVIDIA AI Enterprise streamlines the development and deployment of enterprise-grade accelerated AI software with open-source containers and frameworks optimized for the underlying NVIDIA GPU infrastructure, all with the help of support services .

To learn more, join NVIDIA at Oracle Cloud World in the AI Pavillion, attend this session on the new OCI instances on Wednesday, Sept. 20, and visit these web pages on Oracle Cloud Infrastructure , OCI Compute , how Oracle approaches AI and the NVIDIA AI Platform .

Original URL: https://blogs.nvidia.com/blog/2023/09/19/oracle-cloud-infrastructure-nvidia-gpu-accelerated-compute-instances/