Meta’s Grand Teton Brings NVIDIA Hopper to Its Data Centers

Next-generation AI platform uses NVIDIA H100 GPUs to tackle major AI challenges.

Author: Cliff Edwards

Meta today announced its next-generation AI platform, Grand Teton, including NVIDIA’s collaboration on design.

Compared to the company’s previous generation Zion EX platform, the Grand Teton system packs in more memory, network bandwidth and compute capacity, said Alexis Bjorlin, vice president of Meta Infrastructure Hardware, at the 2022 OCP Global Summit, an Open Compute Project conference.

AI models are used extensively across Facebook for services such as news feed, content recommendations and hate-speech identification, among many other applications.

“We’re excited to showcase this newest family member here at the summit,” Bjorlin said in prepared remarks for the conference, adding her thanks to NVIDIA for its deep collaboration on Grand Teton’s design and continued support of OCP.

Named after the 13,000-foot mountain that crowns one of Wyoming’s two national parks, Grand Teton uses NVIDIA H100 Tensor Core GPUs to train and run AI models that are rapidly growing in their size and capabilities, requiring greater compute.

The NVIDIA Hopper architecture , on which the H100 is based, includes a Transformer Engine to accelerate work on these neural networks, which are often called foundation models because they can address an expanding set of applications from natural language processing to healthcare, robotics and more.

The NVIDIA H100 is designed for performance as well as energy efficiency. H100-accelerated servers, when connected with NVIDIA networking across thousands of servers in hyperscale data centers, can be 300x more energy efficient than CPU-only servers.

“NVIDIA Hopper GPUs are built for solving the world’s tough challenges, delivering accelerated computing with greater energy efficiency and improved performance, while adding scale and lowering costs,” said Ian Buck, vice president of hyperscale and high performance computing at NVIDIA. “With Meta sharing the H100-powered Grand Teton platform, system builders around the world will soon have access to an open design for hyperscale data center compute infrastructure to supercharge AI across industries.”

Grand Teton sports 2x the network bandwidth and 4x the bandwidth between host processors and GPU accelerators compared to Meta’s prior Zion system, Meta said.

The added network bandwidth enables Meta to create larger clusters of systems for training AI models, Bjorlin said. It also packs more memory than Zion to store and run larger AI models.

Packing all these capabilities into one integrated server “dramatically simplifies deployment of systems, allowing us to install and provision our fleet much more rapidly, and increase reliability,” said Bjorlin.

Original URL: https://blogs.nvidia.com/blog/2022/10/18/meta-grand-teton/