HYPERION RESEARCH

HYP_Link

CoreWeave Launches First GB200 Cloud Service for AI

Jaclyn Ludema and Mark Nossokoff
February 2025

RECENT DEVELOPMENT

CoreWeave has announced the general availability of NVIDIA GB200 NVL72-based instances, marking a significant milestone in the cloud computing landscape, particularly for AI and high-performance computing (HPC). CoreWeave claims to be the first cloud provider to offer these advanced instances, which are designed to enhance the training, deployment, and scaling of complex AI models.

Key announced features of interest to the HPC/AI community include:

- **Performance enhancements:** The GB200 NVL72 instances leverage the NVIDIA GB200 Grace Blackwell Superchip, enabling faster real-time large language model (LLM) inferencing compared with previous generations.
- **Cost efficiency:** CoreWeave asserts these instances significantly lower cost of ownership and less energy consumption for real-time inference than prior generations of GPUs.
- **Scalability:** CoreWeave's infrastructure supports clusters of up to 110,000 GPUs using advanced networking technologies, which provide high bandwidth and low latency.

ANALYST COMMENT

Recent Hyperion Research studies have found that organizations are increasingly turning to cloud services to evaluate cutting-edge hardware and software solutions to test feasible use on-premises and/or in the cloud. CoreWeave's announcement highlights an emerging market shift towards specialized GPU as a Service (GPUaaS) options tailored for AI workloads. By standing up this GPUaaS alongside an infrastructure explicitly designed for AI (bare metal compute nodes, high-throughput networking), CoreWeave is seeking to provide an alternative to the typical cloud providers whose infrastructure supports an array of workloads with various priorities. At certain scales, some AI users may find this alternative can save on costs and improve time-to-solution.

The launch of NVIDIA GB200 NVL72-based instances by CoreWeave marks a significant milestone in the HPC/AI market landscape, and other vendors may soon follow suit. By combining cutting-edge GPUaaS with an AI specific infrastructure, CoreWeave is looking to democratize access to advanced AI technologies, potentially encouraging new users/organizations to evaluate AI. This accessibility may lead to a surge in AI-driven applications and solutions across various industries.

Copyright Notice