NVIDIA Ethernet Networking Accelerates World's Largest AI Supercomputer, Built by xAI

NVIDIA Spectrum-X Makes Colossal NVIDIA Hopper 100,000-GPU System Possible

NVIDIA today announced that xAI's Colossus supercomputer cluster comprising 100,000 NVIDIA Hopper GPUs in Memphis, Tennessee, achieved this massive scale by using the NVIDIA Spectrum-X™ Ethernet networking platform, which is designed to deliver superior performance to multi-tenant, hyperscale AI factories using standards-based Ethernet, for its Remote Direct Memory Access (RDMA) network.

Colossus, the world's largest AI supercomputer, is being used to train xAI's Grok family of large language models, with chatbots offered as a feature for X Premium subscribers. xAI is in the process of doubling the size of Colossus to a combined total of 200,000 NVIDIA Hopper GPUs.

The supporting facility and state-of-the-art supercomputer was built by xAI and NVIDIA in just 122 days, instead of the typical timeframe for systems of this size that can take many months to years. It took 19 days from the time the first rack rolled onto the floor until training began.

While training the extremely large Grok model, Colossus achieves unprecedented network performance. Across all three tiers of the network fabric, the system has experienced zero application latency degradation or packet loss due to flow collisions. It has maintained 95% data throughput enabled by Spectrum-X congestion control.

This level of performance cannot be achieved at scale with standard Ethernet, which creates thousands of flow collisions while delivering only 60% data throughput.

"AI is becoming mission-critical and requires increased performance, security, scalability and cost-efficiency," said Gilad Shainer, senior vice president of networking at NVIDIA. "The NVIDIA Spectrum-X Ethernet networking platform is designed to provide innovators such as xAI with faster processing, analysis and execution of AI workloads, and in turn accelerates the development, deployment and time to market of AI solutions."

"Colossus is the most powerful training system in the world," said Elon Musk on X. "Nice work by xAI team, NVIDIA and our many partners/suppliers."

"xAI has built the world's largest, most-powerful supercomputer," said a spokesperson for xAI. "NVIDIA's Hopper GPUs and Spectrum-X allow us to push the boundaries of training AI models at a massive-scale, creating a super-accelerated and optimized AI factory based on the Ethernet standard."

At the heart of the Spectrum-X platform is the Spectrum SN5600 Ethernet switch, which supports port speeds of up to 800Gb/s and is based on the Spectrum-4 switch ASIC. xAI chose to pair the Spectrum-X SN5600 switch with NVIDIA BlueField-3® SuperNICs for unprecedented performance.

Spectrum-X Ethernet networking for AI brings advanced features that deliver highly effective and scalable bandwidth with low latency and short tail latency, previously exclusive to InfiniBand. These features include adaptive routing with NVIDIA Direct Data Placement technology, congestion control, as well as enhanced AI fabric visibility and performance isolation — all key requirements for multi-tenant generative AI clouds and large enterprise environments.

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

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, and performance of NVIDIA's products, services, and technologies, including NVIDIA Hopper GPUs, NVIDIA Spectrum-X Ethernet networking platform, NVIDIA Spectrum SN5600 Ethernet switch, Spectrum-4 switch ASIC, and NVIDIA BlueField-3 SuperNICs; features of xAI's Colossus supercomputer cluster; xAI being in the process of doubling the size of Colossus to a combined total of 200,000 NVIDIA Hopper GPUs; the NVIDIA Spectrum-X Ethernet networking platform being designed to provide innovators such as xAI with faster processing, analysis and execution of AI workloads, and in turn accelerating the development, deployment and time to market of AI solutions; NVIDIA's Hopper GPUs and Spectrum-X allowing xAI to push the boundaries of training AI models at a massive scale, creating a super-accelerated and optimized AI factory based on the Ethernet standard 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.

Alex Shapiro
Enterprise Networking
1-415-608-5044
ashapiro@nvidia.com