NEWS RELEASE

Cisco Reveals Nexus HyperFabric AI Clusters, A New Simplified Data Center Infrastructure Solution with NVIDIA for Generative AI

2024-06-04

News Summary:

- The new solution will combine Cisco and NVIDIA innovation to simplify the deployment of generative AI applications, providing IT visibility and analytics across the entire AI infrastructure stack.
- Cisco Nexus HyperFabric AI clusters make it easy for enterprise customers to build infrastructure to run generative AI models and inference applications without deep IT knowledge and skills.
- Exclusive cloud management capabilities help customers easily deploy, manage and monitor data centers, colocation facilities and edge sites.

LAS VEGAS, June 4, 2024 /PRNewswire/ -- **CISCO LIVE** -- Today Cisco (NASDAQ: CSCO), the leader in security and networking, announced a breakthrough AI cluster solution with NVIDIA for the data center that transforms how customers build, manage and optimize infrastructure and software.

Delivering on the Cisco Networking Cloud vision to simplify networking, Cisco is bringing to market a new enterprise-ready, end-to-end infrastructure solution to scale generative AI workloads. The Cisco Nexus HyperFabric AI cluster solution combines Cisco AI-native networking with NVIDIA accelerated computing and AI software, and a robust VAST data store. It is designed to enable customers to focus on AI-driven innovation and new revenue opportunities rather than IT management.

According to Cisco's recent Global Networking Trends Report in the next two years, 60% of IT leaders and professionals expect to deploy AI-enabled predictive network automation across all domains to better manage NetOps[1]. Additionally, 75% plan to deploy tools that offer end-to-end visibility via a single console into different network domains including campus and branch, WAN, data center, internet, public clouds and industrial networks.

"While the promise of AI is clear, the path forward for many just starting out is not. Customers often face economic and operational challenges to get an AI stack up and running," said Jonathan Davidson, Executive Vice President and General Manager, Cisco Networking. "Cisco is committed to making the deployment and operation of AI infrastructure simpler. Together with NVIDIA, we are delivering a simple-to-deploy, cloud-operated AI-stack solution for on-premises deployments that builds on our Cisco Networking Cloud platform vision for automation and simplicity."

"Generative AI requires purpose-built infrastructure and software that enables enterprises to securely turn their data into fuel for business transformation," said Kevin Deierling, Senior Vice President of Networking at NVIDIA. "NVIDIA and Cisco are providing an enterprise-ready AI platform and control plane to simplify deployment of the accelerated computing, networking and software needed for generative AI workloads."

At Cisco Live, the company is demonstrating how it is committed to helping its customers quickly deploy AI infrastructure. Cisco is also putting the right tools in the hands of its customers to build intuitive AI-native networks, anticipate failures and quickly diagnose and remediate problems.

How Cisco Nexus HyperFabric AI Cluster Works

The on-premises solution features a single place to design, deploy, monitor and assure the AI pods and data center workloads. It guides users from design, to validated deployment, to monitoring and assurance for enterprise-ready AI infrastructure. With its cloud management capabilities, customers can easily deploy and manage large scale fabrics across data centers, colocation facilities and edge sites.

The Cisco Nexus HyperFabric AI cluster solution offers automated, cloud-managed operations across a unified compute and networking fabric combining Cisco's Ethernet switching expertise founded on Cisco Silicon One, integrated with NVIDIA's accelerated computing and NVIDIA AI Enterprise software, and VAST's data storage platform. This will include:

- Cisco cloud management capabilities to simplify IT operations across all phases of the workflow.
- Cisco Nexus 6000 series switches for spine and leaf that deliver 400G and 800G Ethernet fabric performance.
- Cisco Optics family of QSFP-DD modules to offer customer choice and deliver super high densities.
- NVIDIA AI Enterprise software to streamline the development and deployment of production-grade generative AI workloads
- NVIDIA NIM inference microservices that accelerate the deployment of foundation models while ensuring data security, and are available with NVIDIA AI Enterprise
- NVIDIA Tensor Core GPUs starting with the NVIDIA H200 NVL, designed from the ground up to supercharge generative AI workloads with game-changing performance and memory capabilities.
- NVIDIA BlueField-3 data processing unit DPU processor and BlueField-3 SuperNIC for accelerating AI compute networking, data access and security workloads.
- Enterprise reference design for AI built on NVIDIA MGX, a modular and flexible server architecture.
- The VAST Data Platform, which offers unified storage, database and a data-driven function engine built for AI.

Availability

Select customers may have early trial access to this AI solution in Q4 of CY 2024, with general availability expected shortly thereafter.

Introducing AI Skills for Partners and IT Pros

Cisco Learning and Certifications has introduced a new CCDE AI Infrastructure certification with training available now on Cisco U. It provides expert-level Network Engineers and Design Network Engineers with the expertise to translate AI workload business requirements into technical and sustainable best practices for infrastructure design. Cisco has also launched the first stage of its AI partner specializations from Cisco Black Belt. These learning journeys equip Cisco partners with the product knowledge needed to help them build AI practices and accelerate their customer deployments.

Supporting Comments

"The Nexus HyperFabric AI cluster solution is competitively differentiated and leverages Cisco's overall portfolio strength, new cloud-managed subscriptions model, and the Cisco-NVIDIA partnership, among other aspects, for capturing wallet share in Enterprise AI Datacenter Switching and Software Ops solutions."
- *Vijay Bhagavath, Research Vice President, IDC*

"Generative AI models demand lightning-fast access to vast amounts of data, a feat only achieved through unprecedented computing power and high-performance networking infrastructure. With the new Cisco Nexus HyperFabric AI cluster solution, we are helping enterprises build AI data centers with NVIDIA accelerated computing and AI software and Cisco networking, and the VAST Data Platform, providing end-to-end visibility of compute, networking, storage and data management so organizations can seamlessly build and scale their AI operations."
- *Renen Hallak, CEO and co-founder of VAST Data*

"Cisco compute and networking solutions are integral to powering PTC's R&D, data center performance, connectivity, and our Servigistics software to deliver the AI-powered service supply chain capabilities for our customers across industries, helping them better estimate, predict, optimize, and improve service supply chain performance. Cisco's new AI infrastructure solution with NVIDIA opens more possibilities for PTC to scale our data center infrastructure easily and efficiently, enhance operations and performance, and help meet our sustainability goals."
- *Michael Blake, VP of Information Technology at PTC.*

"Ethernet has the broadest scale in our client base, and the Cisco NVIDIA partnership and resulting offers will accelerate our ability to deliver AI solutions to our clients. Both Cisco and NVIDIA are already a significant part of WWT's AI Proving Ground lab where we help clients select and operationalize AI architectures so they can more quickly turn their data into insights and action."
- *Neil Anderson, VP Cloud, Infrastructure, and AI Solutions, WWT*

Added Resources

- Executive Blog: It's Time to Take Control of Experiences and Transform Infrastructure Operations for AI, by Jonathan Davidson
- Executive Blog: AI Takes Center Stage in the Data Center by Kevin Wollenweber and Jeremy Foster
- Press Release: Cisco ThousandEyes Digital Experience Assurance Radically Shifts IT Operations from Monitoring to Action

About Cisco

Cisco (NASDAQ: CSCO) is the worldwide technology leader that securely connects everything to make anything possible. Our purpose is to power an inclusive future for all by helping our customers reimagine their applications, power hybrid work, secure their enterprise, transform their infrastructure, and meet their sustainability goals. Discover more on The Newsroom and follow us on X at @Cisco.

This press release includes a high-level outline of some of Cisco's current product plans. Although it reflects Cisco's current intentions, even the best of plans can change. As such, nothing included in this press release is a binding commitment, and the development, release and timing of any product or feature described is subject to change. Customers should not rely on this information when making a purchase decision, and Cisco will have no liability for delay or failure to deliver any features described.

[1] Cisco Global Networking Trends Report 2024

View original content to download multimedia:https://www.prnewswire.com/news-releases/cisco-reveals-nexus-hyperfabric-ai-clusters-a-new-simplified-data-center-infrastructure-solution-with-nvidia-for-generative-ai-302162908.html

SOURCE Cisco Systems, Inc.