NVIDIA Blackwell Ultra AI Factory Platform Paves Way for Age of AI Reasoning

- **Top Computer Makers, Cloud Service Providers and GPU Cloud Providers to Boost Training and Test-Time Scaling Inference, From Reasoning to Agentic and Physical AI**
- **New Open-Source NVIDIA Dynamo Inference Software to Scale Up Reasoning AI Services With Leaps in Throughput, Faster Response Time and Reduced Total Cost of Ownership**
- **NVIDIA Spectrum-X Enhanced 800G Ethernet Networking for AI Infrastructure Significantly Reduces Latency and Jitter**

GTC—NVIDIA today announced the next evolution of the NVIDIA Blackwell AI factory platform, NVIDIA Blackwell Ultra — paving the way for the age of AI reasoning.

NVIDIA Blackwell Ultra boosts training and test-time scaling inference — the art of applying more compute during inference to improve accuracy — to enable organizations everywhere to accelerate applications such as AI reasoning, agentic AI and physical AI.

Built on the groundbreaking Blackwell architecture introduced a year ago, Blackwell Ultra includes the NVIDIA GB300 NVL72 rack-scale solution and the NVIDIA HGX™ B300 NVL16 system. The GB300 NVL72 delivers 1.5x more AI performance than the NVIDIA GB200 NVL72, as well as increases Blackwell's revenue opportunity by 50x for AI factories, compared with those built with NVIDIA Hopper™.

"AI has made a giant leap — reasoning and agentic AI demand orders of magnitude more computing performance," said Jensen Huang, founder and CEO of NVIDIA. "We designed Blackwell Ultra for this moment — it's a single versatile platform that can easily and efficiently do pretraining, post-training and reasoning AI inference."

NVIDIA Blackwell Ultra Enables AI Reasoning
The NVIDIA GB300 NVL72 connects 72 Blackwell Ultra GPUs and 36 Arm Neoverse-based NVIDIA Grace™ CPUs in a rack-scale design, acting as a single massive GPU built for test-time scaling. With the NVIDIA GB300 NVL72, AI models can access the platform's increased compute capacity to explore different solutions to problems and break down complex requests into multiple steps, resulting in higher-quality responses.

GB300 NVL72 is also expected to be available on NVIDIA DGX™ Cloud, an end-to-end, fully managed AI platform on leading clouds that optimizes performance with software, services and AI expertise for evolving workloads. NVIDIA DGX SuperPOD™ with DGX GB300 systems uses the GB300 NVL72 rack design to provide customers with a turnkey AI factory.

The NVIDIA HGX B300 NVL16 features 11x faster inference on large language models, 7x more compute and 4x larger memory compared with the Hopper generation to deliver breakthrough performance for the most complex workloads, such as AI reasoning.

In addition, the Blackwell Ultra platform is ideal for applications including:

- Agentic AI, which uses sophisticated reasoning and iterative planning to autonomously solve complex, multistep problems. AI agent systems go beyond instruction-following. They can reason, plan and take actions to achieve specific goals.
- Physical AI, enabling companies to generate synthetic, photorealistic videos in real time for the training of applications such as robots and autonomous vehicles at scale.

NVIDIA Scale-Out Infrastructure for Optimal Performance
Advanced scale-out networking is a critical component of AI infrastructure that can deliver top performance while reducing latency and jitter.

Blackwell Ultra systems seamlessly integrate with the NVIDIA Spectrum-X™ Ethernet and NVIDIA Quantum-X800 InfiniBand platforms, with 800 Gb/s of data throughput available for each GPU in the system, through an NVIDIA ConnectX®-8 SuperNIC. This delivers best-in-class remote direct memory access capabilities to enable AI factories and cloud data centers to handle AI reasoning models without bottlenecks.

NVIDIA BlueField®-3 DPUs, also featured in Blackwell Ultra systems, enable multi-tenant networking, GPU compute elasticity, accelerated data access and real-time cybersecurity threat detection.

Global Technology Leaders Embrace Blackwell Ultra
Blackwell Ultra-based products are expected to be available from partners starting from the second half of 2025.

Cisco, Dell Technologies, Hewlett Packard Enterprise, Lenovo and Supermicro are expected to deliver a wide range of servers based on Blackwell Ultra products, in addition to Aivres, ASRock Rack, ASUS, Eviden, Foxconn, GIGABYTE, Inventec, Pegatron, Quanta Cloud Technology (QCT), Wistron and Wiwynn.

Cloud service providers Amazon Web Services, Google Cloud, Microsoft Azure and Oracle Cloud Infrastructure and GPU cloud providers CoreWeave, Crusoe, Lambda, Nebius, Nscale, Yotta and YTL will be among the first to offer Blackwell Ultra-powered instances.

NVIDIA Software Innovations Reduce AI Bottlenecks
The entire NVIDIA Blackwell product portfolio is supported by the full-stack NVIDIA AI platform. The NVIDIA Dynamo open-source inference framework — also announced today — scales up reasoning AI services, delivering leaps in throughput while reducing response times and model serving costs by providing the most efficient solution for scaling test-time compute.

NVIDIA Dynamo is new AI inference-serving software designed to maximize token revenue generation for AI factories deploying reasoning AI models. It orchestrates and accelerates inference communication across thousands of GPUs, and uses disaggregated serving to separate the processing and generation phases of large language models on different GPUs. This allows each phase to be optimized independently for its specific needs and ensures maximum GPU resource utilization.

Blackwell systems are ideal for running new NVIDIA Llama Nemotron Reason models and the NVIDIA AI-Q Blueprint, supported in the NVIDIA AI Enterprise software platform for production-grade AI. NVIDIA AI Enterprise includes NVIDIA NIM™ microservices, as well as AI frameworks, libraries and tools that enterprises can deploy on NVIDIA-accelerated clouds, data centers and workstations.

The Blackwell platform builds on NVIDIA's ecosystem of powerful development tools, NVIDIA CUDA-X™ libraries, over 6 million developers and 4,000+ applications scaling performance across thousands of GPUs.

Learn more by watching the NVIDIA GTC keynote and register for sessions from NVIDIA and industry leaders at the show, which runs through March 21.

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, availability, and performance of NVIDIA's products, services, and technologies; third parties adopting or offering NVIDIA's products and technologies; Blackwell Ultra being able to easily and efficiently do pretraining, post-training and reasoning AI inference; and advanced networking being a critical component of AI infrastructure that can deliver top performance while reducing latency and jitter 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.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein

Kristin Uchiyama
Enterprise and Edge Computing
+1-408-486-2248
kuchiyama@nvidia.com