



NVIDIA Enters Production With Dynamo, the Broadly Adopted Inference Operating System for AI Factories

News Summary:

- NVIDIA Dynamo 1.0 provides a production-grade, open source foundation for inference at scale.
- Dynamo and NVIDIA TensorRT-LLM optimizations integrate natively into open source frameworks such as LangChain, LLM-d, LMCache, SGLang and vLLM to boost inference performance.
- Dynamo boosts inference performance of NVIDIA Blackwell GPUs by up to 7x, lowering token cost and increasing revenue opportunity for millions of GPUs with free, open source software.
- NVIDIA inference platform integrated by cloud service providers, Amazon Web Services (AWS), Microsoft Azure, Google Cloud and Oracle Cloud Infrastructure (OCI), along with NVIDIA cloud partners Alibaba Cloud, CoreWeave, Together AI and Nebius — and adopted by AI-native companies Cursor and Perplexity; inference endpoint providers Baseten, Deep Infra and Fireworks; and global enterprises ByteDance, Meituan, PayPal and Pinterest.

GTC—NVIDIA today announced NVIDIA Dynamo 1.0, open source software for generative and agentic inference at scale, with widespread global adoption. Together with the NVIDIA Blackwell platform, Dynamo 1.0 enables cloud providers, AI innovators and global enterprises to deliver high-performance AI inference with unmatched scale, efficiency and speed.

As agentic AI systems move into production across industries, scaling inference within a data center has become a complex challenge of resource orchestration, with requests of varying sizes and modalities, as well as performance objectives, arriving in unpredictable bursts.

Just as a computer's operating system coordinates hardware and applications, Dynamo 1.0 functions as the distributed "operating system" of AI factories, seamlessly orchestrating GPU and memory resources across the cluster to power complex AI workloads. In recent industry benchmarks, Dynamo boosted the inference performance of NVIDIA Blackwell GPUs by up to 7x, lowering token cost and increasing revenue opportunity for millions of GPUs with free, open source software.

"Inference is the engine of intelligence, powering every query, every agent and every application," said Jensen Huang, founder and CEO of NVIDIA. "With NVIDIA Dynamo, we've created the first-ever 'operating system' for AI factories. The rapid adoption across our ecosystem shows this next wave of agentic AI is here, and NVIDIA is powering it at global scale."

Dynamo 1.0 splits inference work across GPUs by adding smarter "traffic control" and the ability to move data between GPUs and lower-cost storage, reducing wasted work and easing memory limits. For agentic AI and long prompts, it can route requests to GPUs that already have the most relevant "short-term memory" from earlier steps, then offload that memory when it is not needed.

NVIDIA Inference Platform Gains Momentum

NVIDIA is accelerating the open source ecosystem by integrating Dynamo and NVIDIA TensorRT™-LLM library optimizations into popular frameworks from providers such as LangChain, LLM-d, LMCache, SGLang, vLLM and more. Core Dynamo building blocks like KVBM for smarter memory management, NVIDIA NIXL for fast GPU-to-GPU data movement and NVIDIA Grove for simplified scaling are also available as standalone modules. NVIDIA also contributes TensorRT-LLM CUDA® kernels to the FlashInfer project so they can be natively integrated into open source frameworks.

The NVIDIA inference platform is supported across the AI ecosystem, including:

- **Cloud Service Providers:** [Amazon Web Services](#) (AWS), [Microsoft Azure](#), [Google Cloud](#), [OCI](#)
- **NVIDIA Cloud Partners:** [Alibaba Cloud](#), CoreWeave, Crusoe, DigitalOcean, Gcore, GMI Cloud, Lightning AI, Nebius, Nscale, Together AI, Vultr
- **AI-Native Companies:** Cursor, Hebbia, Perplexity
- **Inference Endpoint Providers:** Baseten, Deep Infra, Fireworks
- **Global Enterprises:** AstraZeneca, BlackRock, ByteDance, Coupang, Instacart, Meituan, PayPal, Pinterest, Shopee, SoftBank Corp.

Chen Goldberg, executive vice president of product and engineering at CoreWeave, said: "As AI moves from experimental pilots to continuous, large-scale production, the underlying infrastructure must be as dynamic as the models it supports. Supporting NVIDIA Dynamo allows us to offer a more seamless, resilient environment for deploying complex AI agents. This foundation provides the durability and high-performance orchestration required to move the industry's most ambitious agentic workloads into global production."

Danila Shtan, chief technology officer of Nebius, said: "Delivering reliable AI inference at scale isn't just about powerful GPUs, it's about the software that turns that performance into real customer outcomes. We value how NVIDIA's software

stack, from Dynamo to TensorRT-LLM, brings deep optimization, predictable performance and faster time to deployment, helping us offer customers a simpler, higher-performance path to production AI.”

Matt Madrigal, chief technology officer of Pinterest, said: “Delivering an intuitive, multimodal AI experience to hundreds of millions of users requires real-time intelligence at global scale. As a significant adopter in open source, we’re committed to building scalable AI technologies. With NVIDIA Dynamo optimizing our deployment, we’re expanding the seamless and personalized experiences we deliver, powered by high-performance AI infrastructure.”

Vipul Ved Prakash, cofounder and CEO of Together AI, said: “AI natives require inference that can reliably and efficiently scale with their application. NVIDIA Dynamo 1.0, combined with cutting-edge inference research from Together AI, helps us deliver a high-performance stack to offer accelerated, cost-effective inference for large-scale production workloads.”

Dynamo 1.0 is available today to developers worldwide. To learn more and get started, read the [blog](#) and visit the [Dynamo webpage](#).

About NVIDIA

[NVIDIA](#) (NASDAQ: NVDA) is the world leader in AI and accelerated computing.

Certain statements in this press release including, but not limited to, statements as to: Inference being the engine of intelligence, powering every query, every agent and every application; NVIDIA powering the next wave of agentic AI at global scale; the benefits, impact, performance, and availability of NVIDIA's products, services, and technologies; expectations with respect to NVIDIA's third party arrangements, including with its collaborators and partners; expectations with respect to technology developments; and other statements that are not historical facts are forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, which are subject to the “safe harbor” created by those sections based on management’s beliefs and assumptions and on information currently available to management and 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 and political conditions; NVIDIA's reliance on third parties to manufacture, assemble, package and test NVIDIA's products; the impact of technological development and competition; development of new products and technologies or enhancements to NVIDIA's existing product and technologies; market acceptance of NVIDIA's products or NVIDIA's partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of NVIDIA's products or technologies when integrated into systems; and changes in applicable laws and regulations, 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.

© 2026 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA, NVIDIA Hopper and TensorRT are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

Jordan Byrnes
NVIDIA
press@nvidia.com