



NVIDIA Dynamo Open-Source Library Accelerates and Scales AI Reasoning Models

NVIDIA Dynamo Increases Inference Performance While Lowering Costs for Scaling Test-Time Compute; Inference Optimizations on NVIDIA Blackwell Boosts Throughput by 30x on DeepSeek-R1

GTC—NVIDIA today unveiled [NVIDIA Dynamo](#), an open-source inference software for accelerating and scaling AI reasoning models in AI factories at the lowest cost and with the highest efficiency.

Efficiently orchestrating and coordinating AI inference requests across a large fleet of GPUs is crucial to ensuring that AI factories run at the lowest possible cost to maximize token revenue generation.

As AI reasoning goes mainstream, every AI model will generate tens of thousands of [tokens](#) used to “think” with every prompt. Increasing inference performance while continually lowering the cost of inference accelerates growth and boosts revenue opportunities for service providers.

NVIDIA Dynamo, the successor to NVIDIA Triton Inference Server™, 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 (LLMs) on different GPUs. This allows each phase to be optimized independently for its specific needs and ensures maximum GPU resource utilization.

“Industries around the world are training AI models to think and learn in different ways, making them more sophisticated over time,” said Jensen Huang, founder and CEO of NVIDIA. “To enable a future of custom reasoning AI, NVIDIA Dynamo helps serve these models at scale, driving cost savings and efficiencies across AI factories.”

Using the same number of GPUs, Dynamo doubles the performance and revenue of AI factories serving Llama models on today’s NVIDIA Hopper™ platform. When running the DeepSeek-R1 model on a large cluster of GB200 NVL72 racks, NVIDIA Dynamo’s intelligent inference optimizations also boost the number of tokens generated by over 30x per GPU.

To achieve these inference performance improvements, NVIDIA Dynamo incorporates features that enable it to increase throughput and reduce costs. It can dynamically add, remove and reallocate GPUs in response to fluctuating request volumes and types, as well as pinpoint specific GPUs in large clusters that can minimize response computations and route queries. It can also offload inference data to more affordable memory and storage devices and quickly retrieve them when needed, minimizing inference costs.

NVIDIA Dynamo is fully open source and supports PyTorch, SGLang, NVIDIA TensorRT™-LLM and vLLM to allow enterprises, startups and researchers to develop and optimize ways to serve AI models across disaggregated inference. It will enable users to accelerate the adoption of AI inference, including at AWS, Cohere, CoreWeave, Dell, Fireworks, Google Cloud, Lambda, Meta, Microsoft Azure, Nebius, NetApp, OCI, Perplexity, Together AI and VAST.

Inference Supercharged

NVIDIA Dynamo maps the knowledge that inference systems hold in memory from serving prior requests — known as KV cache — across potentially thousands of GPUs.

It then routes new inference requests to the GPUs that have the best knowledge match, avoiding costly recomputations and freeing up GPUs to respond to new incoming requests.

“To handle hundreds of millions of requests monthly, we rely on NVIDIA GPUs and inference software to deliver the performance, reliability and scale our business and users demand,” said Denis Yarats, chief technology officer of Perplexity AI. “We look forward to leveraging Dynamo, with its enhanced distributed serving capabilities, to drive even more inference-serving efficiencies and meet the compute demands of new AI reasoning models.”

Agentic AI

AI provider Cohere is planning to power agentic AI capabilities in its Command series of models using NVIDIA Dynamo.

“Scaling advanced AI models requires sophisticated multi-GPU scheduling, seamless coordination and low-latency communication libraries that transfer reasoning contexts seamlessly across memory and storage,” said Saurabh Baji, senior vice president of engineering at Cohere. “We expect NVIDIA Dynamo will help us deliver a premier user experience to our enterprise customers.”

Disaggregated Serving

The NVIDIA Dynamo inference platform also supports disaggregated serving, which assigns the different computational phases of LLMs — including building an understanding of the user query and then generating the best response — to different GPUs. This approach is ideal for reasoning models like the [new NVIDIA Llama Nemotron model family](#), which uses advanced inference techniques for improved contextual understanding and response generation. Disaggregated serving allows each phase to be fine-tuned and resourced independently, improving throughput and delivering faster responses to users.

Together AI, the AI Acceleration Cloud, is looking to integrate its proprietary Together Inference Engine with NVIDIA Dynamo to enable seamless scaling of inference workloads across GPU nodes. This also lets Together AI dynamically address traffic bottlenecks at various stages of the model pipeline.

“Scaling reasoning models cost effectively requires new advanced inference techniques, including disaggregated serving and context-aware routing,” said Ce Zhang, chief technology officer of Together AI. “Together AI provides industry-leading performance using our proprietary inference engine. The openness and modularity of NVIDIA Dynamo will allow us to seamlessly plug its components into our engine to serve more requests while optimizing resource utilization — maximizing our accelerated computing investment. We’re excited to leverage the platform’s breakthrough capabilities to cost-effectively bring open-source reasoning models to our users.”

NVIDIA Dynamo Unpacked

NVIDIA Dynamo includes four key innovations that reduce inference serving costs and improve user experience:

- **GPU Planner:** A planning engine that dynamically adds and removes GPUs to adjust to fluctuating user demand, avoiding GPU over- or under-provisioning.
- **Smart Router:** An LLM-aware router that directs requests across large GPU fleets to minimize costly GPU recomputations of repeat or overlapping requests — freeing up GPUs to respond to new incoming requests.
- **Low-Latency Communication Library:** An inference-optimized library that supports state-of-the-art GPU-to-GPU communication and abstracts complexity of data exchange across heterogeneous devices, accelerating data transfer.
- **Memory Manager:** An engine that intelligently offloads and reloads inference data to and from lower-cost memory and storage devices without impacting user experience.

NVIDIA Dynamo will be made available in NVIDIA NIM™ microservices and supported in a future release by the NVIDIA AI Enterprise software platform with production-grade security, support and stability.

Learn more by watching the [NVIDIA GTC keynote](#), reading this [blog on Dynamo](#) and [registering 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 NVIDIA’s products and technologies and the benefits and impact thereof; industries around the world training AI models to think and learn in different ways, making them more sophisticated over time; and to enable a future of custom reasoning AI, NVIDIA Dynamo helping serve these models at scale, driving cost savings and efficiencies across AI factories 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.

© 2025 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NVIDIA Hopper, NVIDIA NIM, NVIDIA Triton Inference Server 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.

Cliff Edwards
NVIDIA Corporation
+1-415-699-2755
cliffe@nvidia.com