



# NVIDIA Groq 3 LPX Now in Full Production With World-Class Speed for Agentic AI

## Groq 3 LPX Extends the Vera Rubin Platform, Delivering Ultrafast Token Generation for the Next Generation of Agentic AI, With Nebius the First to Adopt

### News Summary:

- In Artificial Analysis benchmarking, Groq 3 LPX showcased world-class speed for agentic coding and other latency-sensitive workloads.
- NVIDIA Groq 3 LPX extends the inference performance of NVIDIA Vera Rubin NVL72 systems by dramatically increasing token generation rates.
- Nebius is the first AI cloud to adopt NVIDIA Groq 3 LPX.

**Hot Chips**--NVIDIA today announced that [NVIDIA Groq 3 LPX](#), the interactive AI inference accelerator, is now in full production. An extension of the NVIDIA Vera Rubin platform, Groq 3 LPX delivers a major boost in AI inference by enabling ultrafast token generation for highly responsive agentic systems.

Agentic systems can generate massive volumes of tokens across hundreds or thousands of inference steps, making faster token generation critical for agents to reason, act and complete complex tasks in real time.

Vera Rubin NVL72 systems provide the most versatile training and inference platform for every AI factory. NVIDIA Groq 3 LPX extends the inference performance of Vera Rubin NVL72 by dramatically increasing the rate of token generation, providing premium user experiences for context-heavy workloads so agents can act at extreme speeds.

NVIDIA Groq 3 LPX is pushing the frontier of AI inference. It delivered a record 3,400 output tokens per second in Artificial Analysis benchmarking running Gemma 4 31B, an open source agentic model, with a 100,000-token context critical for agentic systems -- the fastest performance ever recorded for the model.

Groq 3 LPX enables agentic tasks such as coding in minutes versus hours, providing 4x faster responsiveness for agents and latency-sensitive workloads than the nearest alternative platform.

"Inference is the growth engine of AI. NVIDIA Grace Blackwell and NVL72 revolutionized large language model inference with an unprecedented leap in performance and efficiency," said Jensen Huang, founder and CEO of NVIDIA. "Vera Rubin extends that vision with workload-optimized AI factory configurations designed for the era of agentic AI, advancing the performance frontier with LPX for ultrafast token generation. This transforms how intelligence is produced, delivering another giant leap in AI throughput, efficiency and responsiveness, just as demand for AI computation is accelerating worldwide."

### Groq 3 LPX -- The Interactive AI Inference Accelerator

Agentic AI creates two distinct computing challenges: efficiently processing enormous amounts of context and generating tokens with extremely low latency.

NVIDIA Groq 3 LPX is purpose-built to extend Vera Rubin's interactivity -- the rate at which tokens are generated for an individual user, determining how quickly an agent can complete each step of its work.

Faster generation gives agents more time to inspect files, write and test code, call tools, verify results and iterate while maintaining a responsive user experience.

### AI Cloud Momentum for Groq 3 LPX

AI clouds are becoming the engines of the AI economy, giving enterprises and developers access to advanced infrastructure for training, reasoning and inference at scale. For providers serving latency-sensitive, high-volume inference workloads, NVIDIA Groq 3 LPX provides a path to deploy differentiated compute in proven rack-scale systems.

[Nebius](#), a leading AI cloud, plans to bring NVIDIA Groq 3 LPX to [Nebius Token Factory](#), its production inference platform, giving developers access to extreme token generation speed for highly responsive agentic AI applications.

"Generation is the phase of inference that determines how responsive an AI system actually is, and that's exactly what NVIDIA Groq 3 LPX is built to accelerate," said Danila Shtan, chief technology officer of Nebius. "As the first AI cloud bringing it to production via Nebius Token Factory, we're making sure every step of an agent's loop feels instant -- through the same API developers are already using, with no migration to a new stack."

Following Nebius, purpose-built AI inference cloud [Groq](#) plans to be among the platform's earliest adopters.

## **Extreme Codesign for AI Factories**

Through extreme codesign across seven chips and five purpose-built racks, NVIDIA Vera Rubin is the most extensive AI factory platform.

NVIDIA Vera Rubin NVL72 and Groq 3 LPX tackle the various workload requirements of customer AI factories, including frontier model makers and open model service providers.

These rack platforms feature NVIDIA BlueField®-4 DPUs and work in combination with NVIDIA Vera CPU racks, NVIDIA Vera BlueField-4 STX storage and NVIDIA Spectrum™-6 SPX Ethernet to optimize multi-agent systems for the highest throughput per watt and the lowest-latency inference.

## **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: expectations with respect to growth, performance, availability, and benefits of NVIDIA's products, services and technologies, and related trends and drivers; expectations with respect to NVIDIA's third party arrangements, including with its collaborators and partners; expectations with respect to technology developments, and related trends and drivers; projected market growth and trends; expectations with respect to AI and related industries; 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 products 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; NVIDIA's ability to realize the potential benefits of business investments or acquisitions; 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 NVIDIA'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.

© 2026 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, BlueField and NVIDIA Spectrum are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Groq and LPU are used under license from Groq, Inc. 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.

Alex Shapiro  
Corporate Communications  
NVIDIA Corporation  
[press@nvidia.com](mailto:press@nvidia.com)