



NVIDIA Unveils Rubin CPX: A New Class of GPU Designed for Massive-Context Inference

News Summary:

- The NVIDIA Rubin CPX GPU is purpose-built to handle million-token coding and generative video applications.
- The NVIDIA Vera Rubin NVL144 CPX platform packs 8 exaflops of AI performance and 100TB of fast memory in a single rack.
- Companies can monetize at an unprecedented scale, with \$5B in token revenue for every \$100M invested.
- AI innovators like Cursor, Runway and Magic are exploring how Rubin CPX can accelerate their applications.

AI Infra Summit—NVIDIA® today announced NVIDIA Rubin CPX, a new class of GPU purpose-built for massive-context processing. This enables AI systems to handle million-token software coding and generative video with groundbreaking speed and efficiency.

Rubin CPX works hand in hand with NVIDIA Vera CPUs and Rubin GPUs inside the new NVIDIA Vera Rubin NVL144 CPX platform. This integrated NVIDIA MGX system packs 8 exaflops of AI compute to provide 7.5x more AI performance than NVIDIA GB300 NVL72 systems, as well as 100TB of fast memory and 1.7 petabytes per second of memory bandwidth in a single rack. A dedicated Rubin CPX compute tray will also be offered for customers looking to reuse existing Vera Rubin NVL144 systems.

“The Vera Rubin platform will mark another leap in the frontier of AI computing — introducing both the next-generation Rubin GPU and a new category of processors called CPX,” said Jensen Huang, founder and CEO of NVIDIA. “Just as RTX revolutionized graphics and physical AI, Rubin CPX is the first CUDA GPU purpose-built for massive-context AI, where models reason across millions of tokens of knowledge at once.”

[NVIDIA Rubin CPX](#) enables the highest performance and token revenue for long-context processing — far beyond what today’s systems were designed to handle. This transforms AI coding assistants from simple code-generation tools into sophisticated systems that can comprehend and optimize large-scale software projects.

To process video, AI models can take up to 1 million tokens for an hour of content, pushing the limits of traditional GPU compute. Rubin CPX integrates video decoder and encoders, as well as long-context inference processing, in a single chip for unprecedented capabilities in long-format applications such as video search and high-quality generative video.

Built on the NVIDIA Rubin architecture, the Rubin CPX GPU uses a cost-efficient, monolithic die design packed with powerful NVFP4 computing resources and is optimized to deliver extremely high performance and energy efficiency for AI inference tasks.

Advancements Offered by Rubin CPX

Rubin CPX delivers up to 30 petaflops of compute with NVFP4 precision for the highest performance and accuracy. It features 128GB of cost-efficient GDDR7 memory to accelerate the most demanding context-based workloads. In addition, it delivers 3x faster attention capabilities compared with NVIDIA GB300 NVL72 systems — boosting an AI model’s ability to process longer context sequences without a drop in speed.

Rubin CPX is offered in multiple configurations, including the Vera Rubin NVL144 CPX, that can be combined with the [NVIDIA Quantum-X800 InfiniBand](#) scale-out compute fabric or the [NVIDIA Spectrum-X™ Ethernet](#) networking platform with [NVIDIA Spectrum-XGS Ethernet](#) technology and NVIDIA ConnectX®-9 SuperNICs™. Vera Rubin NVL144 CPX enables companies to monetize at an unprecedented scale, with \$5 billion in token revenue for every \$100 million invested.

Industry Leaders Look to Rubin CPX

AI innovators are exploring how Rubin CPX can accelerate their applications, ranging from large-scale software development to the analysis of dynamic visual content to better understand moving images.

Cursor, an AI-powered software company that offers an advanced code editor, sees the benefits of Rubin CPX to boost developer productivity with intelligent code generation and collaborative tools directly in the coding environment.

“With NVIDIA Rubin CPX, Cursor will be able to deliver lightning-fast code generation and developer insights, transforming software creation,” said Michael Truell, CEO of Cursor. “This will unlock new levels of productivity and empower users to ship ideas once out of reach.”

Runway, an American generative AI company, will use NVIDIA technologies to enable creators to produce cinematic content and sophisticated visual effects with unmatched scale and efficiency.

“Video generation is rapidly advancing toward longer context and more flexible, agent-driven creative workflows,” said Cristóbal Valenzuela, CEO of Runway. “We see Rubin CPX as a major leap in performance, supporting these demanding workloads to build more general, intelligent creative tools. This means creators — from independent artists to major studios — can gain unprecedented speed, realism and control in their work.”

Magic is an AI research and product company developing foundation models to power AI agents that can automate software engineering.

“With a 100-million-token context window, our models can see a codebase, years of interaction history, documentation and libraries in context without fine-tuning,” said Eric Steinberger, CEO of Magic. “This enables users to coach the agent at test time through conversation and access to their environments, bringing us closer to autonomous agentic experiences. Using a GPU like NVIDIA Rubin CPX greatly accelerates our compute workloads.”

Software Support

NVIDIA Rubin CPX will be supported by the complete NVIDIA AI stack — from accelerated infrastructure to enterprise-ready software. The [NVIDIA Dynamo](#) platform efficiently scales AI inference, dramatically boosting throughput while cutting response times and model serving costs.

The processors will be able to run the latest in the NVIDIA Nemotron™ family of multimodal models that provide state-of-the-art reasoning for enterprise-ready AI agents. For production-grade AI, Nemotron models can be delivered with NVIDIA AI Enterprise, a software platform that includes [NVIDIA NIM](#)™ microservices as well as AI frameworks, libraries and tools that enterprises can deploy on NVIDIA-accelerated clouds, data centers and workstations.

Built on decades of innovation, the Rubin platform extends NVIDIA’s developer ecosystem — with [NVIDIA CUDA-X](#)™ libraries, a community of over 6 million developers and nearly 6,000 CUDA applications.

Availability

NVIDIA Rubin CPX is expected to be available at the end of 2026.

Learn more by watching NVIDIA Vice President of Hyperscale and High-Performance Computing Ian Buck’s [keynote at AI Infra Summit](#) on Sept. 9 at 10am PT.

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: Vera Rubin systems continuing to deliver extraordinary performance and efficiency; with Rubin CPX, building a GPU uniquely suited for million-token context processing, cutting the cost of inference and unlocking advanced capabilities for developers and creators everywhere; 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.

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 and all other NVIDIA trademarks mentioned herein 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.

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