

09 Sep 2025 18:24:29 ET | 11 pages

NVIDIA Corp (NVDA.O)

AI Infra Summit Takeaways

CITI'S TAKE

Today, NVIDIA's VP of Hyperscale and High-Performance Computing, Ian Buck, reiterated the company's goal to accelerate Gen AI adoption through GPU-powered data centers at AI Infrastructure Conference in Santa Clara, California. While various discussions emerged during the keynote, we highlight the following two: **1) The new Rubin CPX GPU** which when combined with Vera and Rubin can deliver up to \$5B in token revenue for every \$100M invested (~50x ROI vs GB200 NVL72's ~10x). **2) NVDA announced its GB300 NVL72 rack-scale system set a new reasoning inference benchmark records vs. GB200 NVL72** (see body for details). We note by inserting the Rubin CPX into the roadmap NVDA is accelerating its one-year cadence amid the ramping ASIC competition (see our [recent note](#)). We also see these announcements as an indication that the age of inference is upon us as reflected by Google's recent comments of 50x+ token increase Y/Y.

Buy

Price (09 Sep 25 16:00)	US\$170.76
Target price	US\$200.00
Expected share price return	17.1%
Expected dividend yield	0.0%
Expected total return	17.1%
Market Cap	US\$4,149,468M

Atif Malik^{AC}

Papa Sylla

NVDA announced the following at AI Infra Summit:

[Rubin CPX: A New Class of GPU Designed for Long-Context Inference](#): NVIDIA Rubin CPX is a new class of GPU purpose-built geared for the highest performance and token revenue for ultra-large context processing. Rubin CPX enables AI systems to handle million-token software coding and generative video with groundbreaking speed and efficiency. Rubin CPX delivers up to 3x faster attention capabilities compared with NVIDIA GB300 NVL72 systems. Per NVDA, the chip works tightly with NVIDIA Vera CPUs and Rubin GPUs inside the new NVIDIA Vera Rubin NVL144 CPX platform, which enables companies to monetize investments at an unprecedented scale with ~50x ROI (\$5B in token revenue for every \$100M invested).

See Appendix A-1 for Analyst Certification, Important Disclosures and Research Analyst Affiliations

Citi Research is a division of Citigroup Global Markets Inc. (the "Firm"), which does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the Firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision. Certain products (not inconsistent with the author's published research) are available only on Citi's portals.

The diagram illustrates the VR NVL144 CPX Compute Tray. On the left, a vertical server rack is shown with a white box highlighting a specific compute tray. A dashed line connects this box to a detailed view of the tray on the right. The detailed view shows the internal components of the compute tray, including the Rubin Compute Tray, Vera, Rubin CPX, and ConnectX-9.

VR NVL144 CPX
Compute Tray

Rubin

Vera

Rubin CPX

ConnectX-9

Source: Nvidia

NVIDIA Drives Continuous Innovation With One-Year Rhythm
Full-stack | One architecture | CUDA everywhere

	Blackwell	Rubin	Feynman
COMPUTE	<ul style="list-style-type: none"> RTX 5090 RTX 5080 RTX 5070 RTX 5060 	<ul style="list-style-type: none"> RTX 6090 RTX 6080 RTX 6070 RTX 6060 	<ul style="list-style-type: none"> RTX 7090 RTX 7080 RTX 7070 RTX 7060
SCALE (SCALE-GPU)	<ul style="list-style-type: none"> Grace CPU Grace Hopper Superchip 	<ul style="list-style-type: none"> Grace CPU Grace Hopper Superchip 	<ul style="list-style-type: none"> Grace CPU Grace Hopper Superchip
NETWORK (SCALE-IO)	<ul style="list-style-type: none"> BlueField-3 SoC BlueField-3 SoC 	<ul style="list-style-type: none"> BlueField-3 SoC BlueField-3 SoC 	<ul style="list-style-type: none"> BlueField-3 SoC BlueField-3 SoC
SYSTEM	<ul style="list-style-type: none"> Blackwell Superchip Blackwell Superchip 	<ul style="list-style-type: none"> Rubin Superchip Rubin Superchip 	<ul style="list-style-type: none"> Feynman Superchip Feynman Superchip

Timeline: 2025, 2026, 2027, 2028

NVIDIA

© 2025 Citigroup Inc. No redistribution without Citigroup's written permission.

Source: Citi Research, Nvidia's AI Infra Summit Keynote

NVIDIA Blackwell Ultra Sets the Bar in New MLPerf Inference Benchmark: NVDA announced that its NVIDIA GB300 NVL72 rack-scale system set records on the new reasoning inference benchmark in the latest MLPerf inference benchmark. NVDA's newest system can deliver up to 1.4x more DeepSeek-R1 inference throughput compared to GB200 NVL72 systems. The company added that the platform also set performance records on all new data center benchmarks added to the MLPerf Inference v5.1 suite — including DeepSeek-R1, Llama 3.1 405B Interactive, Llama 3.1 8B and Whisper. These system level performance are on top of NVDA's per-GPU records across MLPerf data center benchmark.

NVIDIA Corp

Valuation

Our price target for NVDA of \$200 is based on ~30x P/E on C26E EPS. Our 30x P/E multiple is in-line with 3-5 year average.

Risks

Downside risks to the attainment of our target price include: 1) competition on gaming could drive the stock lower if Nvidia loses market share; 2) slower-than-expected adoption of new platforms can drive lower data center and gaming sales; 3) lumpiness in auto and data center markets can add volatility to the stock/multiple; and 4) cryptomining impact on gaming sales.

If you are visually impaired and would like to speak to a Citi representative regarding the details of the graphics in this document, please call USA 1-888-500-5008 (TTY: 711), from outside the US +1-210-677-3788