NVIDIA CES 2026 Impact Report

Benchmarks & Strategic Analysis

Executive Summary
CES 2026 marked a pivotal shift for NVIDIA, moving beyond 'Just LLMs' to 'Physical AI' and 'Agentic AI'. With the introduction of the Vera Rubin platform, NVIDIA has quintupled AI performance (50 PFLOPS) and nearly tripled memory bandwidth (22 TB/s) to solve the 'Memory Wall' for next-gen reasoning models.

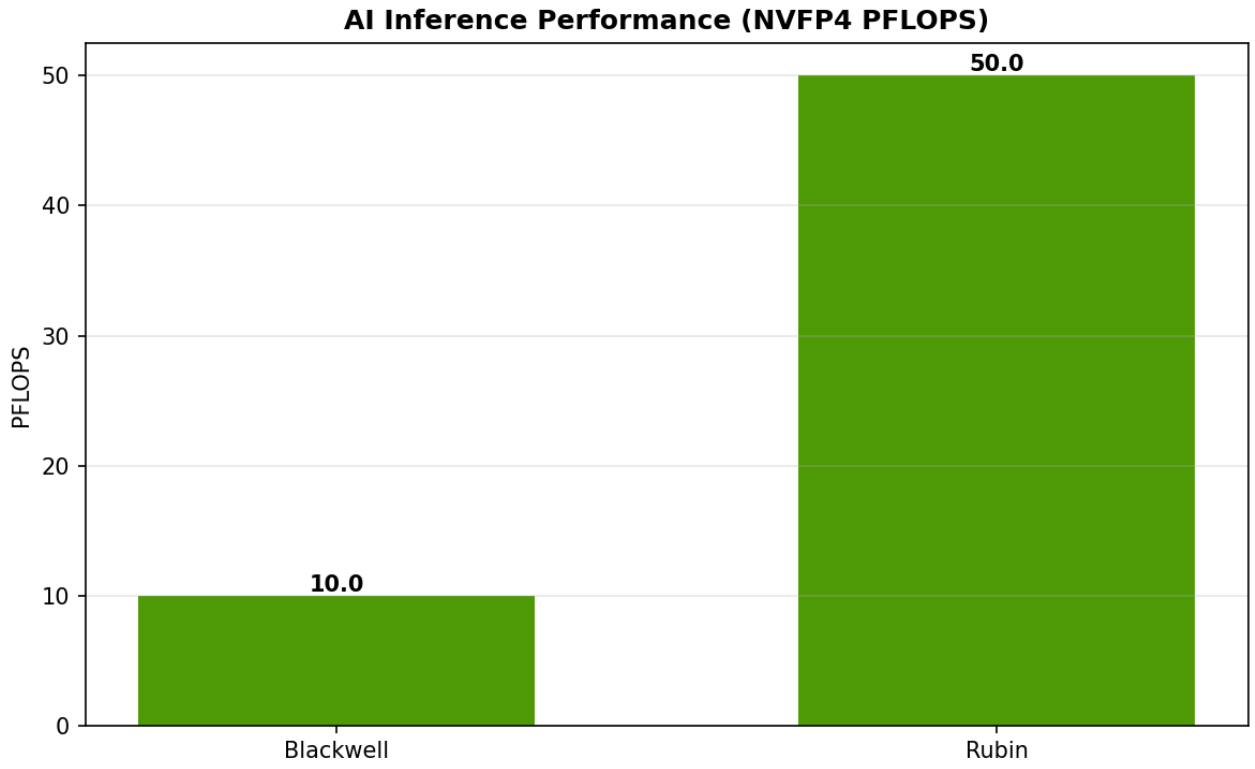
1. The Vera Rubin Platform (Technical Specs)

1. Vera Rubin Platform: The Quantifiable Leap

NVIDIA Vera CPU
- Cores: 88 Custom Olympus Cores (176 Threads)
- Memory: 1.5 TB LPDDR5X System Memory
- Improvement: 3X Memory Capacity vs. Grace CPU
- Transistors: 227 Billion

NVIDIA Rubin GPU
- NVFP4 Inference: 50 PFLOPS (5X vs Blackwell)
- NVFP4 Training: 35 PFLOPS (3.5X vs Blackwell)
- HBM4 Bandwidth: 22 TB/s (2.8X vs Blackwell)
- NVLink Bandwidth: 3.6 TB/s (2X vs Blackwell)
- Transistors: 336 Billion (1.6X vs Blackwell)

Vera Rubin NVL72 (Rack Scale)
- Total Inference: 3.6 ExaFLOPS (5X improvement)
- Total Training: 2.5 ExaFLOPS (3.5X improvement)
- Memory Capacity: 54 TB LPDDR5X (3X) + 20.7 TB HBM (1.5X)
- Scale-Up Bandwidth: 260 TB/s (2X improvement)



Rubin delivers a 5x leap in inference performance over Blackwell, specifically optimized for 'Thinking' models (Test-Time Scaling).

2. The Economics of Intelligence

Insane Demand for AI Computing
Key scaling laws driving the industry (Source: EPOCH AI & Artificial Analysis): 1. Model Size: Growing 10x Parameters per Year. 2. Test-Time Scaling ('Thinking'): Demanding 5x more Tokens per Year. 3. Cost Efficiency: Token Cost dropping 10x Cheaper per Year.

Open Model Ecosystem
NVIDIA is empowering an open ecosystem including: - Major Models: DeepSeek, Qwen (Alibaba), Kimi, Llama (Meta), Mistral, Google. - Strategic Focus: 'Physical AI' (Cosmos) and 'AV' (Alpamayo) are the next frontiers.

3. Infrastructure & Networking Revolution

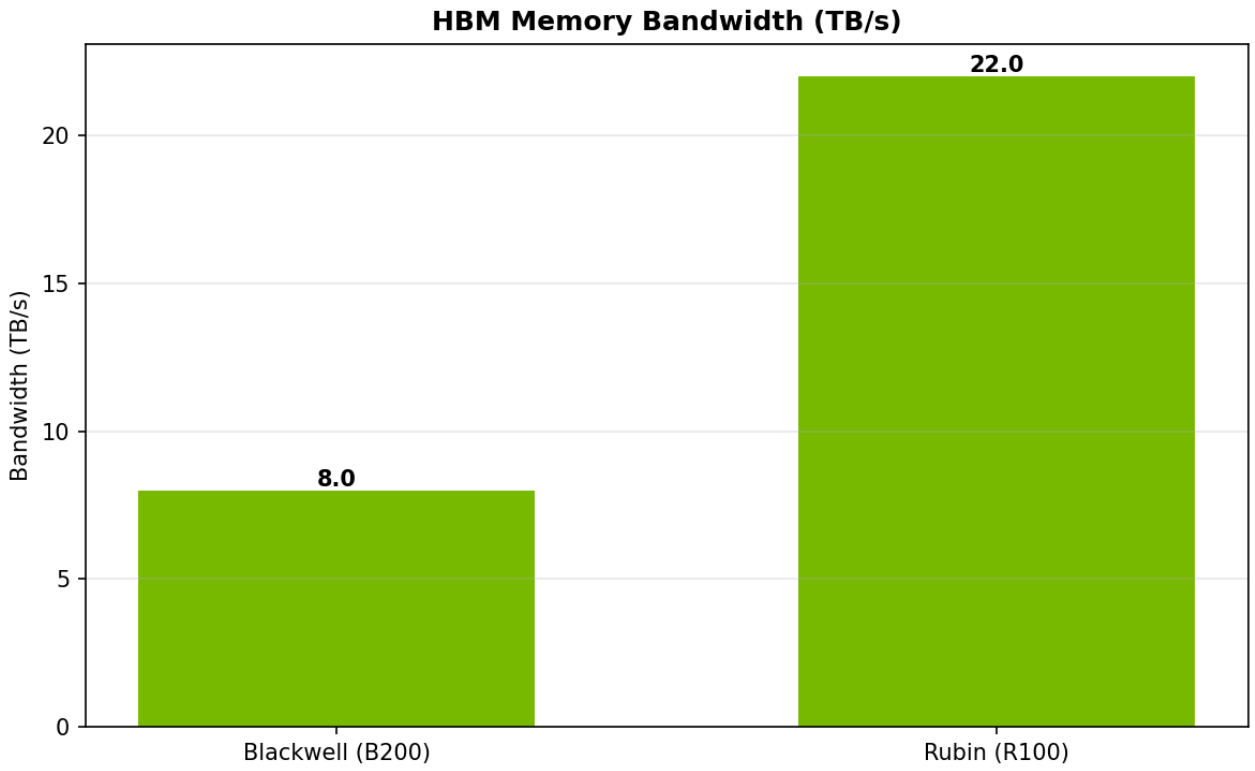
Networking: The Nervous System
- ConnectX-9 SuperNIC: 800 Gb/s Ethernet, Programmable RDMA.
- BlueField-4 DPU: 800 Gb/s, 6x Compute vs BF3, 64-Core Grace CPU embedded.
- NVLink 6 Switch: 3.6 TB/s per-GPU bandwidth, enabling the NVL72 rack to act as one giant GPU.
- Spectrum-X: Ethernet Co-Packaged Optics, scaling to 512 ports of 200 Gb/s.

4. Physical AI & Industrial Digital Twins

Strategic Partnerships (EDA/CAE)
NVIDIA is integrating with the world's leading engineering platforms: - EDA Partners: Cadence (Cerebrus, Virtuoso), Synopsys (AgentEngineer), Siemens. - CAE/Simulation: Ansys, Beta CAE. - Use Cases: Semiconductor DT, Industrial DT, Robotics DT, Automotive DT.

5. Healthcare: AI Learns Laws of Nature

BioMedical Revolution
- BioNeMo Platform: Open-source models (Proteina, RNAPro) for drug discovery.
- Agentic AI: 'Hirable' agents for clinical trials and patient care.
- Lab Agents: Thermo Fisher Scientific building AI-powered lab infrastructure.
- Impact: The $4.9T Healthcare industry is adopting AI at 3x the speed of the U.S. economy.



HBM4 (22 TB/s) is the enabler for these 'Physical AI' workloads which require massive real-time data ingestion.

3. CES 2026 Strategic Highlights

Key Themes: Physical & Agentic AI
1. Physical AI & Robotics: - 'Robotics is the next wave of AI'. - Alpamayo: Open Reasoning VLA for Autonomous Vehicles. - GR00T: Foundation model for humanoid robots. 2. Agentic AI & Cosmos: - Agents are now Multi-Model, Multi-Cloud, and Hybrid. - 'Compute is Data' -> NVIDIA Cosmos platform. 3. New Infrastructure: - ConnectX-9 SuperNIC & Spectrum-X for Ethernet Scale-Out. - 'Context Memory Storage' platform using BlueField-4.

4. System Architecture & Flow

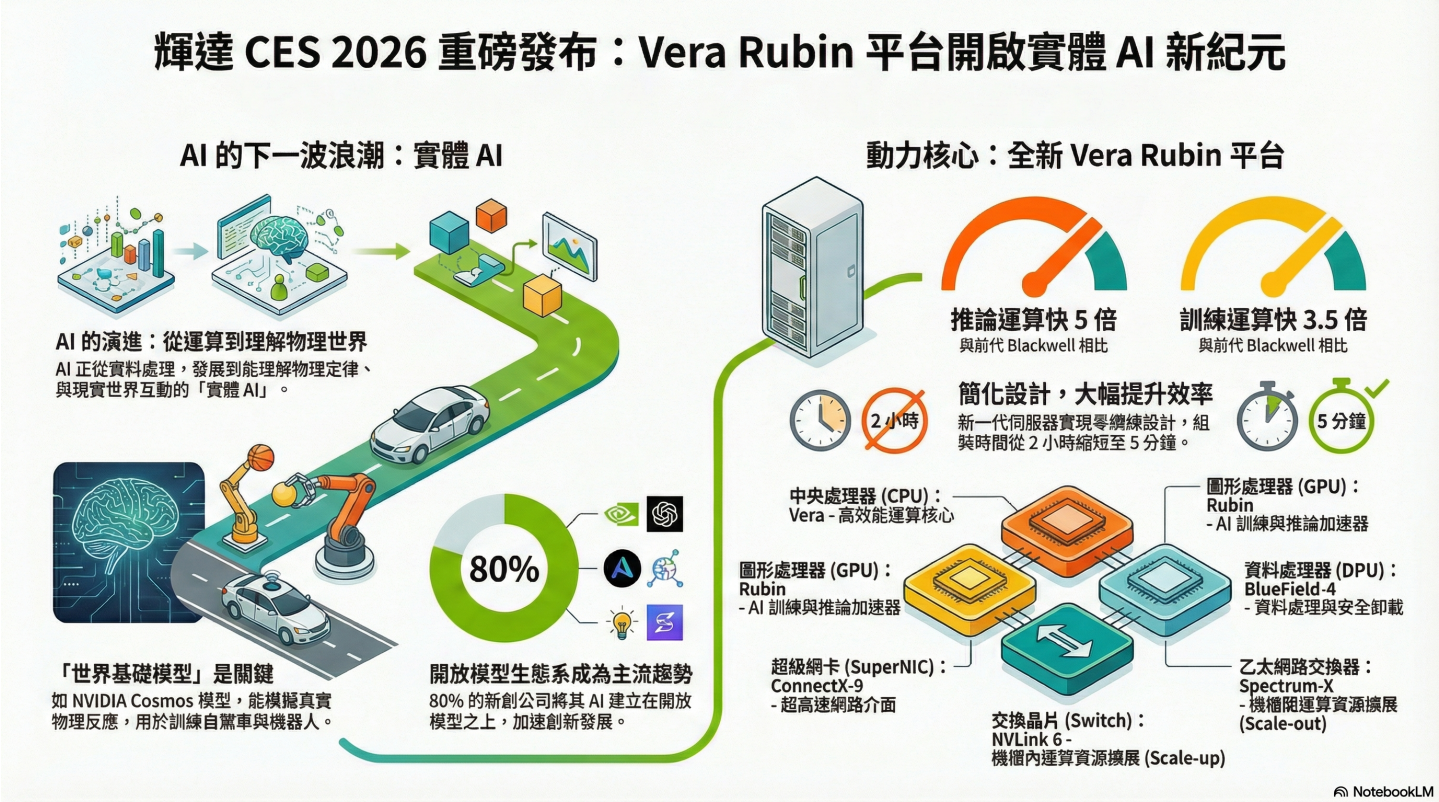


Fig 1. NVIDIA AI Flow & Ecosystem Visualization (Source: CES 2026)