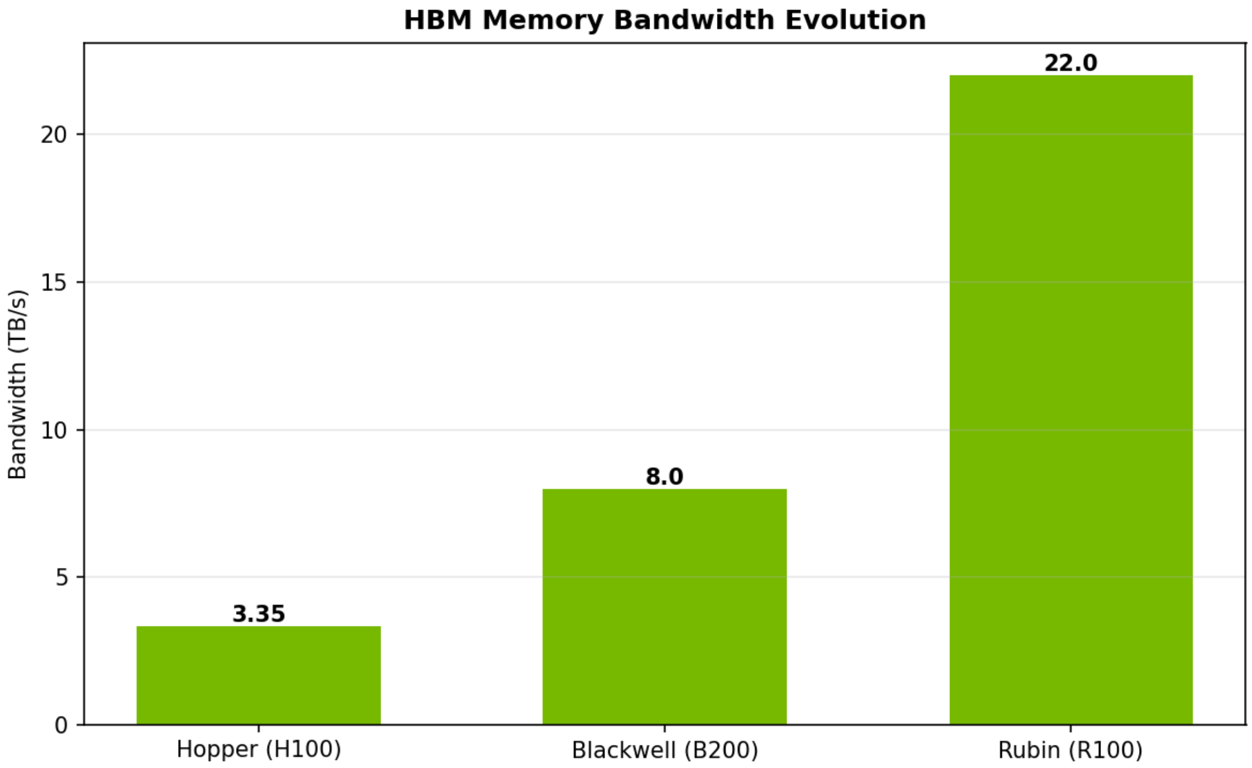
NVIDIA 2026 Platform Benchmark

Rubin vs. Blackwell Generation Gap

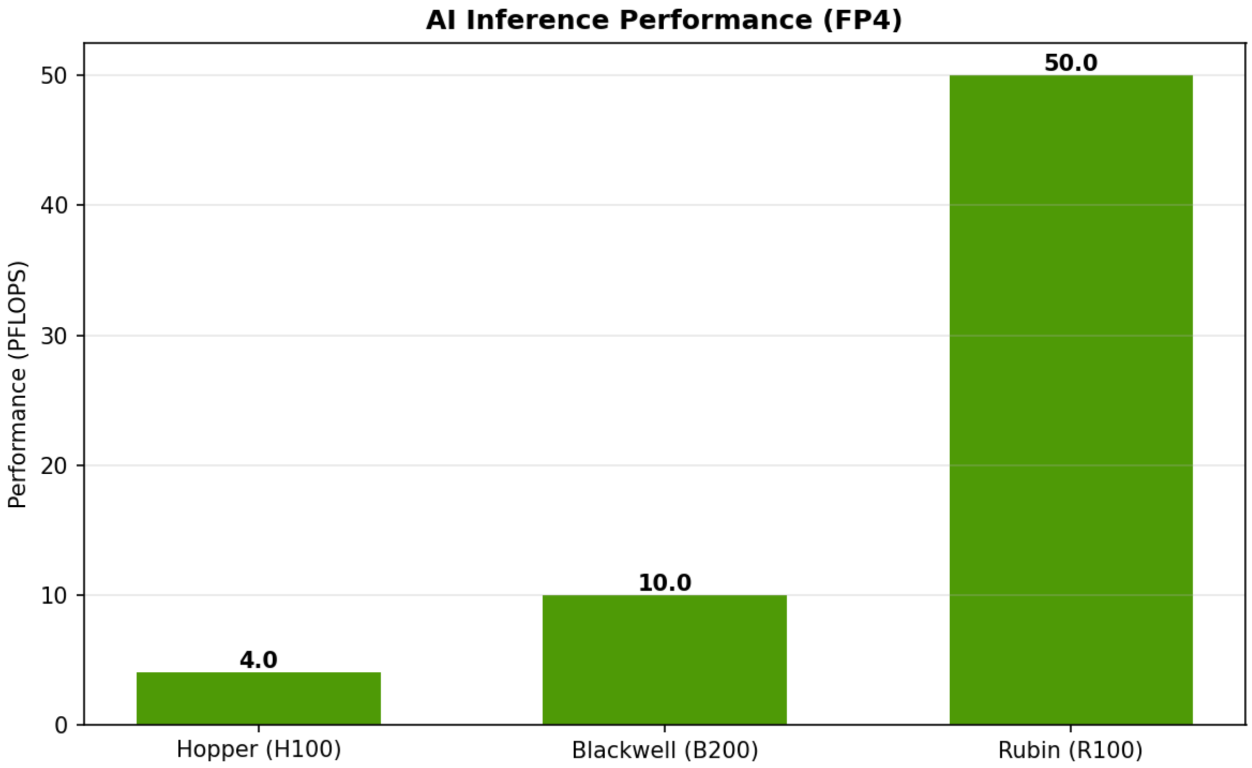
Executive Summary
The 2026 NVIDIA Rubin platform represents a massive leap over the Blackwell generation. Driven by specific bottlenecks in Agentic AI and MoE (Mixture of Experts) models, Rubin triples the memory bandwidth and quintuples inference performance.

1. Memory Bandwidth Revolution (HBM4)



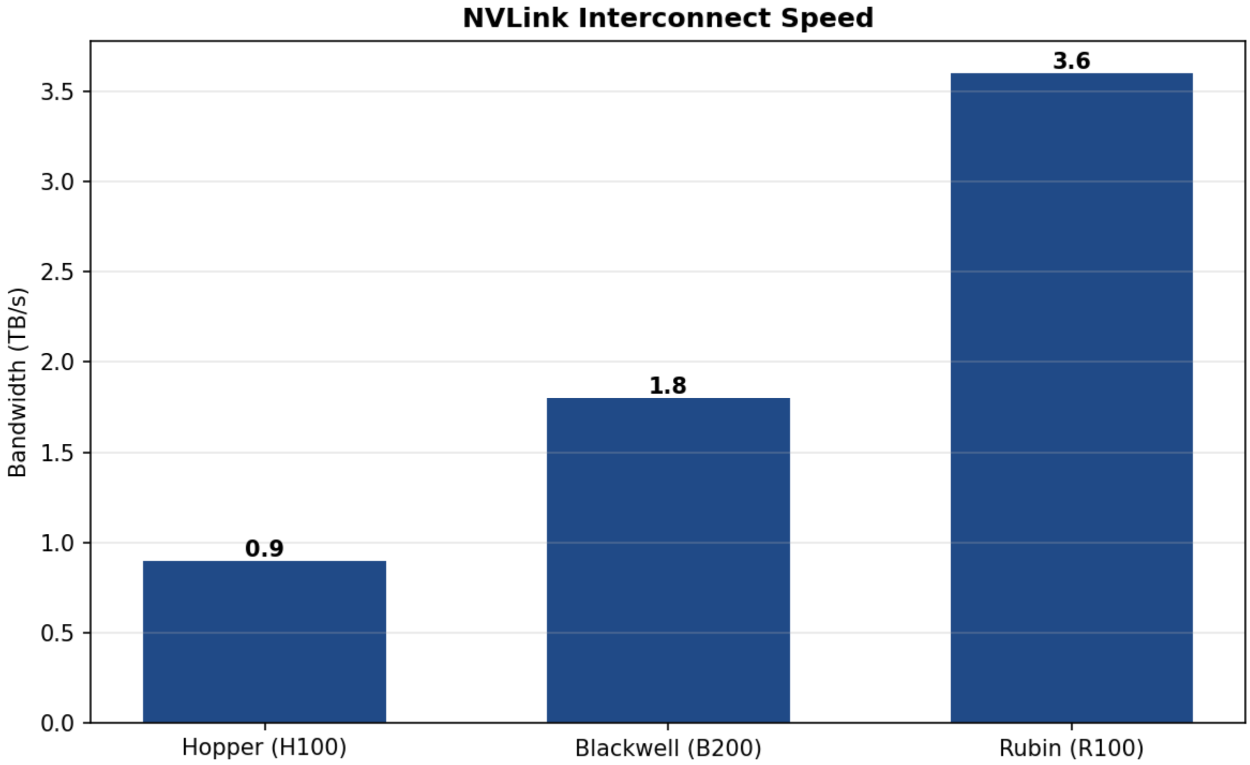
Analysis: Solving the Bottleneck
Rubin introduces HBM4 memory, achieving a staggering 22 TB/s bandwidth per chip. This is nearly 3x the bandwidth of Blackwell (8 TB/s). Why it matters: Large Language Models (LLMs) are 'memory bandwidth bound' during the decoding phase. A 3x increase in bandwidth translates directly to a ~3x increase in tokens-per-second for single-user generation, or massive concurrency gains for serving millions of users.

2. Inference Performance (FP4)



Analysis: Powering Agentic AI
Rubin delivers ~50 PFLOPS of dense AI performance (NVFP4), a 5x jump from Blackwell. This is achieved through the 3nm process node and architectural improvements. This enables 'Physical AI' and complex reasoning agents that require massive compute per token.

3. NVLink 6 Interconnect



Analysis: The Super-Chip Era
NVLink 6 doubles the chip-to-chip speed to 3.6 TB/s. This is critical for the 'Vera Rubin NVL72' rack, allowing 72 GPUs to act as a single giant GPU with unified memory. It minimizes latency when models are split across multiple chips (Tensor Parallelism).

Conclusion

Strategic Outlook
The shift from Blackwell to Rubin is not just an incremental update; it is a structural change specifically designed for the next phase of AI: Agents and Robotics. With 22 TB/s bandwidth and 50 PFLOPS compute, Rubin solves the 'memory wall' that currently limits long-context and reasoning-heavy models.