perplexity

From AI infra perspective: Why NVIDIA’s $5B bet on Intel is bigger than fabs

▯ The fact:
NVIDIA just put $5 billion into Intel (~4% stake) and announced co-development of custom x86 CPUs and PC/AI platforms
(press release).

This is more than a manufacturing hedge — it’s about bridging the deepest gap in AI infrastructure.

▯ My thinking:

1▯ GPU vs CPU memory fabrics
NVLink meshes GPU memory.
RDMA/InfiniBand (and now Enfabrica’s superNICs) mesh CPU memory.
What separates them? PCIe. And Intel CPUs own the strongest PCIe root complex (RC), with cache-coherence built in. That makes Intel the natural bridge between GPU and CPU fabrics.

2▯ From NVLink CC to PCIe
In GB200/GB300 Arm-based systems, NVIDIA uses NVLink CC (cache coherent) to unify CPU+GPU. But overlaying that onto PCIe with RDMA cards isn’t straightforward on Arm.
The recent 48-lane x8 PCIe design in GB systems is powerful — but not as elegant as NVIDIA envisioned. Intel’s x86 IP and coherence experience are the missing link — just as Mellanox once brought RDMA and collectives into the NVLink ecosystem.

3▯ NVIDIA was a chipset company
NVIDIA started life as a chipset player. After its legal settlement with Intel, it effectively stepped away from the x86 chipset business.
NVLink thrives in Arm systems — but in x86, NVIDIA may need to return to its origin. By partnering with Intel, it could be preparing to reboot its chipset role inside x86 systems.
The big question: what role will NVLink logic play inside future x86 CPUs?

What to expect next?
• x86-based GB-style combo systems — inference monsters, leveraging x86 memory systems + PCIe cache coherence.
• Potential convergence of CXL + NVLink CC — a new, cache-coherent “PCIe++” fabric for the AI era.

NVIDIA isn't just hedging fabs. It's reshaping the memory backbone of x86 AI systems.

i need to create confluence blog on this.. i want folks to understand GPU and CPU and LLM to this current merger

NVIDIA's $5 billion investment in Intel marks a strategic, infrastructure-defining move that goes far beyond manufacturing: it's a bet on reshaping the memory backbone of x86 AI systems and bridging core gaps between GPU and CPU architectures critical to the future of large language models (LLMs) and AI workloads.[1] [2] [3]

The Deal: A Brief Overview

- NVIDIA invested $5 billion for ~4% of Intel, signaling a multi-generation R&D partnership.[4] [1]
- Together, they will co-develop custom x86 CPUs for AI data centers and PC platforms that tightly integrate NVIDIA's GPU and AI acceleration technologies with Intel's memory and interconnect expertise.[2] [3] [5]
- The collaboration is not just about custom chips—it's about fundamentally transforming data movement, cache coherence, and AI scaling across x86-centric enterprise infrastructure.[6] [3]

The GPU vs CPU Memory Fabric Divide

- Modern AI workloads depend on massive, low-latency data movement between GPU and CPU memory.[7] [6]
- NVIDIA's NVLink interconnect enables ultra-high-bandwidth GPU-to-GPU and GPU-to-CPU communication—far exceeding PCIe bandwidth, with up to 1.8 TB/s per GPU in GB300-based systems.[7]
- However, x86 systems have long been limited by PCIe's bandwidth and lack of direct cache coherence, while Arm-based NVIDIA systems leverage NVLink CC (cache coherent) and RDMA technologies for unified memory fabrics.[8] [6]
- Intel's PCIe root complex and cache coherence algorithms are the backbone of enterprise-grade memory fabric, providing the "natural bridge" for merging GPU and CPU ecosystems at rack scale.[3] [2]

Why Intel? Why Now?

- NVIDIA historically exited x86 chipsets after its legal settlement with Intel, focusing on GPUs and Arm-based Grace CPUs—where NVLink can operate at full capability.[3]
- By re-entering the x86 ecosystem—jointly designing CPUs with integrated NVLink logic—NVIDIA can bypass existing PCIe bottlenecks and tap Intel's expertise in cache-coherent fabrics, advanced packaging (EMIB/Foveros), and memory architecture.[2] [6]
- This facilitates true rack-scale AI systems (e.g., dozens of CPUs and GPUs acting as a unified memory and compute pool), essential for trillion-parameter LLMs and hyperscale

inference pipelines.[3] [7]

Expected Innovations: x86 GB-Style Combo Systems

- Upcoming “Intel x86 RTX SoCs” and custom x86 data center CPUs will fuse NVIDIA GPUs and AI accelerators directly onto Intel’s CPU die or interconnect, leveraging chiplet and modular design.[6] [2] [3]
- Expect the convergence of NVLink CC, CXL (Compute Express Link), and new PCIe++ innovations—enabling cache-coherent, high-bandwidth fabrics unifying CPU and GPU memory for AI workloads.[2] [3]
- Key outcomes:
  - “Inference monsters”: x86 servers integrating NVLink directly for massive AI acceleration.[2] [3]
  - New opportunities for scaling LLMs and generative AI across enterprise and hyperscale clouds still dominated by x86, with fewer architectural bottlenecks.[6] [3]

Why It Matters for LLMs and AI Infrastructure

- Training and deploying large language models require not just powerful GPUs, but seamless CPU-to-GPU memory coherency, ultra-low latency, and high bandwidth for model hosting and distributed inference.[7] [3]
- This partnership enables:
  - Accelerated model training by eliminating PCIe bottlenecks.
  - Simplified software/hardware integration for AI enterprises built on x86.
  - The groundwork for hybrid architectures where LLMs run on tightly coupled CPU+GPU clusters at datacenter scale.[6] [2]

Takeaway: Beyond a Fab Hedge

- NVIDIA’s move is less about process technology and foundry risk mitigation, and more about orchestrating the merger of the world’s dominant software ecosystem (x86) with the fastest AI hardware stack (NVLink+GPU).[3] [2]
- The result: a new, cache-coherent, memory-rich infrastructure for the AI era, transforming how LLMs and generative models are operated in real-world datacenters.[7] [6]

This deal is where the future of AI takes shape—not only at the chip and server level—but across every level of how enterprise and hyperscale clouds build, scale, and deploy LLMs.[2] [3]

For your Confluence blog, use this structure to walk readers through:

- “AI infra memory fabrics: the CPU-GPU gap”
- “How NVIDIA+Intel are rebuilding the backbone”
- “What this means for x86 AI, LLMs, and every enterprise platform”

⁂

1. https://finance.yahoo.com/news/nvidia-invests-5-billion-intel-130611670.html
2. https://www.tomshardware.com/pc-components/cpus/nvidia-and-intel-announce-jointly-developed-intel-x86-rtx-socs-for-pcs-with-nvidia-graphics-also-custom-nvidia-data-center-x86-processors-nvidia-buys-usd5-billion-in-intel-stock-in-seismic-deal
3. https://www.rdworldonline.com/nvidia-becomes-major-intel-cpu-buyer-in-5b-collaboration/
4. https://www.euronews.com/business/2025/09/18/nvidia-invests-5-billion-in-intel-to-develop-ai-infrastructure
5. https://www.eetimes.com/nvidia-puts-5bn-into-intel-and-will-jointly-develop-ai-superchips/
6. https://www.ai-buzz.com/intel-cpu-with-nvidia-nvlink-creates-unified-ai-platform
7. https://www.nvidia.com/en-us/data-center/nvlink/
8. https://www.nvidia.com/en-us/data-center/nvlink-c2c/
9. http://nvidianews.nvidia.com/news/nvidia-and-intel-to-develop-ai-infrastructure-and-personal-computing-products
10. https://www.axios.com/2025/09/18/ai-nvidia-intel
11. https://www.reuters.com/world/asia-pacific/nvidia-bets-big-intel-with-5-billion-stake-chip-partnership-2025-09-18/
12. https://www.pbs.org/newshour/economy/nvidia-to-invest-5-billion-in-intel-companies-will-work-together-on-ai-infrastructure-and-pcs
13. https://www.tomshardware.com/pc-components/cpus/teams-at-nvidia-and-intel-have-been-working-in-secret-on-jointly-developed-processors-for-a-year-the-trump-administration-has-no-involvement-in-this-partnership-at-all
14. https://www.cnbc.com/2025/09/18/nvidias-huang-calls-5-billion-intel-stake-an-incredible-investment.html
15. https://www.techrepublic.com/article/news-nvidia-intel-deal/
16. https://www.windowscentral.com/hardware/intel/intel-and-nvidia-announce-partnership-that-will-see-jointly-developed-x86-intel-cpus-fused-with-rtx-gpus-in-shocking-move
17. https://www.youtube.com/watch?v=eysh528MbgU
18. https://investor.nvidia.com/news/press-release-details/2024/Computer-Industry-Joins-NVIDIA-to-Build-AI-Factories-and-Data-Centers-for-the-Next-Industrial-Revolution/default.aspx
19. https://folsomtimes.com/nvidia-invests-5b-in-intel-as-tech-leaders-join-forces-on-ai-future/
20. https://apnews.com/article/nvida-intel-chips-investment-73c307d2f6ceccd6854d6666775358f3