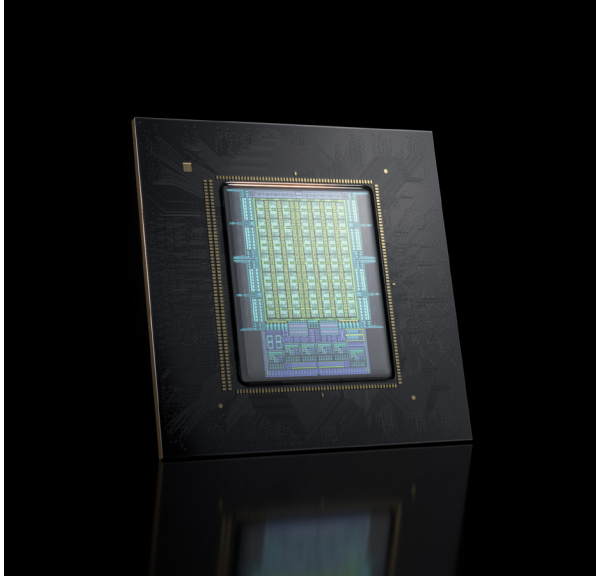
Datasheet

NVIDIA Vera CPU

The CPU for agents.



NVIDIA Vera CPU

AI Scaling Enters the Agentic Inference Era

AI is moving from generating answers to completing work. Agentic AI turns inference into a loop: models generate tokens, code, queries, and tool calls, while CPUs execute those actions, return results, and supply context for the next step.

That loop changes the role of the CPU in the AI factory. Tool calls, sandboxed code execution, data queries, retrieval, evaluation, and orchestration now sit on the critical path—determining how quickly agents can take their next action and how fast training systems can return new feedback.

AI factory economics are shifting from cores per dollar to tokens per dollar—how much useful AI work the infrastructure can produce. As models call more tools, use more data, and run more environments, CPUs become essential to keeping GPUs, agents, and data pipelines moving with less waiting.

NVIDIA Vera was built for this new compute pattern. With fast, efficient Olympus CPU cores, high memory bandwidth, and a uniform data movement architecture, NVIDIA Vera accelerates the CPU-bound work around the model—from agentic sandbox execution and data analytics to orchestration and GPU hosting—helping AI factories deliver more work per watt and per dollar.

Key Offerings

> NVIDIA Vera CPU

> NVIDIA Vera CPU Rack

NVIDIA Vera CPU Architectural Innovations

NVIDIA Vera brings together the core, memory, fabric, GPU interconnect, security, and software ecosystem required to run agentic AI at AI-factory scale. Each element is designed to reduce latency, increase throughput, and keep AI factory infrastructure moving through repeated cycles of reasoning, action, observation, and evaluation.

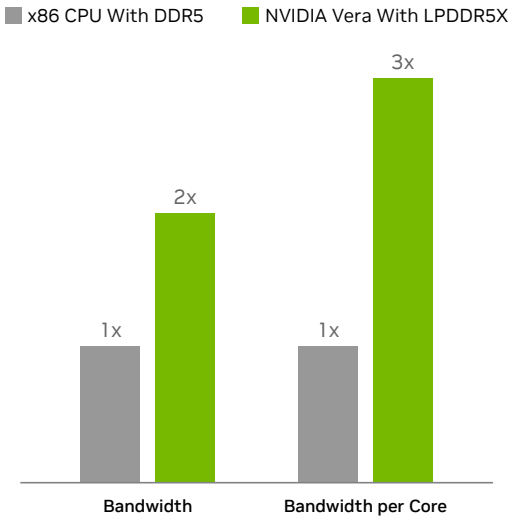
Olympus Cores With NVIDIA Spatial Multithreading

At the heart of NVIDIA Vera are 88 NVIDIA-designed Olympus cores built for the sequential, branch-heavy work that defines agentic AI. Python runtimes, sandboxed code, compilers, queries, and tool chains often sit on the critical path between model steps, so faster per-core execution directly reduces agent, user, and GPU

wait time. To help with this, Vera Olympus cores deliver up to 50% higher IPC than NVIDIA Grace™, while NVIDIA Spatial Multithreading exposes two hardware threads per core—176 per CPU—partitioning key resources to sustain utilization with more predictable performance under dense sandbox load.

LPDDR5X Memory and Second-Generation NVIDIA SCF on a Monolithic Die

NVIDIA Vera uses a single monolithic compute die, connecting all 88 Olympus cores, the unified 164 MB L3 cache, memory, and I/O through the second-generation NVIDIA Scalable Coherency Fabric. The 3.4 TB/s coherent on-chip mesh avoids cross-die latency cliffs and NUMA tuning, helping sustain predictable performance when all cores are active. NVIDIA Vera's LPDDR5X memory subsystem delivers up to 1.2 TB/s of bandwidth and up to 1.5 TB of capacity, with 3x higher bandwidth per core and 40% lower loaded peak latency than traditional CPUs. Using SOCAMM modules, NVIDIA Vera delivers roughly 2x the bandwidth at half the memory power, keeping retrieval, analytics, and sandbox execution fed more efficiently.



Relative performance based on measured data, subject to change. NVIDIA Vera CPU with LPDDR5X performance baselined to x86 CPU with DDR5 across key CPU memory performance metrics.

NVLink-C2C, Confidential Computing, and Day-One Arm Software

NVIDIA Vera extends the CPU into the accelerated AI platform, helping move data, manage memory, and coordinate system control across training, post-training, and inference. Paired with NVIDIA Rubin GPUs through second-generation NVIDIA NVLink™-C2C, NVIDIA Vera is the only CPU coherently connected to NVIDIA Rubin, providing up to 1.8 TB/s of CPU-GPU bandwidth without the bottlenecks of a PCIe-only connection.

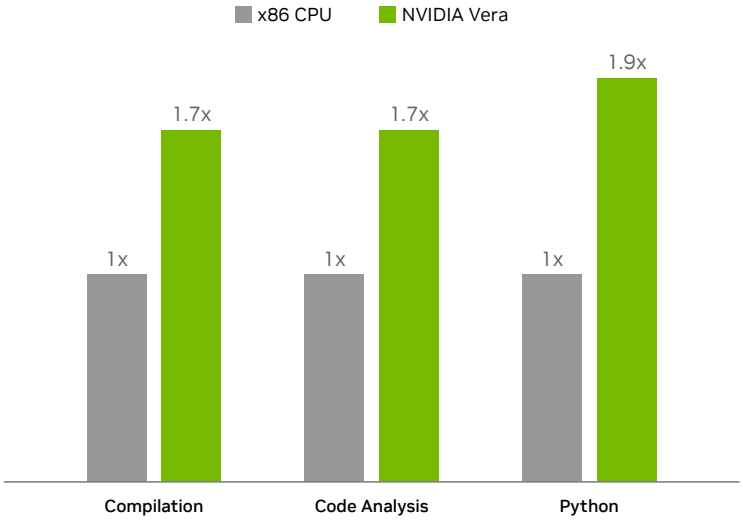
This coherent CPU-GPU architecture gives applications unified access across CPU and GPU memory for KV-cache offload, multi-model execution, and data staging. NVIDIA Vera is fully Armv9.2 compatible, supports NVIDIA Confidential Computing for trusted execution environments across single-CPU and rack-scale deployments and runs major Linux distributions, NVIDIA NGC™ containers, GCC, LLVM, NVIDIA compilers, and the full NVIDIA software stack from day one.

NVIDIA Vera CPU Rack: Built for Agentic AI at Factory Scale

In agentic AI and reinforcement learning, faster sandboxes mean faster evaluation loops, more useful training signal, and less waiting around the model. The NVIDIA Vera CPU Rack packages that capability at factory scale, integrating 256 liquid-cooled NVIDIA Vera CPUs in an NVIDIA MGX™ modular architecture. NVIDIA Vera CPU Rack helps AI factories add dense, deployable CPU capacity alongside GPU infrastructure—accelerating time to market while improving performance per rack for agentic inference and reinforcement learning.

Industry-Leading Agentic CPU Performance

Agentic AI depends on fast CPU execution around the model. Agents compile generated code, run Python tool chains, analyze outputs, and query data before the next model step can begin. NVIDIA Vera accelerates these critical sandbox workloads by up 1.8x versus leading x86 CPUs, speeding the agentic inner loop and helping AI factories increase output per rack, per watt, and per dollar.



Relative performance based on measured data, subject to change. NVIDIA Vera CPU performance baselined to x86 CPU across a variety of workloads, including code compilation, code analysis, and Python.

Solutions for Every Data Center

NVIDIA Vera is designed to scale across every AI factory footprint, from dense liquid-cooled racks to flexible air-cooled servers. The NVIDIA Vera CPU Rack delivers maximum CPU density for large-scale agentic AI and reinforcement learning, while air-cooled two- and single-socket systems bring Vera to enterprise, cloud, and data centers using the same architecture and software stack to enterprise and cloud deployments.

Why Vera Wins for Agentic AI

- **Architected for the agentic loop:** Vera was designed from day one for sandbox execution, tool calls, RL feedback and orchestration.
- **Single monolithic compute die:** All 88 cores share a unified 164 MB L3 cache and 3.4 TB/s on-chip fabric.
- **NVIDIA Spatial Multithreading:** Physically partitions core resources, eliminating the thread-contention slowdowns under load.
- Unmatched memory bandwidth-per-core: Up to 14 GB/s of LPDDR5X bandwidth per core, keeping cores fed.
- **Coherent CPU-GPU unified memory:** NVIDIA NVLink-C2C delivers 1.8 TB/s of coherent bandwidth between Vera and NVIDIA Rubin GPUs or between two CPUs in a two-socket system.

Technical Specifications[1]

	NVIDIA Vera CPU	NVIDIA Vera CPU Rack
Configuration	1 NVIDIA Vera CPU	256 NVIDIA Vera CPUs
Cores | Threads	88 custom NVIDIA Olympus cores (Arm compatible) | 176 threads	22,528 custom NVIDIA Olympus cores (Arm compatible, 88 per CPU) | 45,056 threads (176 per CPU)
L2 Cache (per core)	2 MB	2 MB
Unified L3 Cache	164 MB	42 GB (164 MB per CPU)
SIMD (per core)	6x 128bSVE2 FP8	6x 128bSVE2 FP8
Memory Capacity	Up to 1.5 TB SOCAMM LPDDR5X	Up to 400 TB[2] SOCAMM LPDDR5X
Peak Memory Bandwidth	Up to 1.2 TB/s	Up to 300 TB/s aggregate
NVLINK-C2C Bandwidth	1.8 TB/s	1.8 TB/s per CPU
PCIe CXL	88 PCIe Gen 6 (CPU-only) 96 PCIe Gen 6 (Vera Rubin) x16, x8, x4, x2 bifurcation CXL 3.1	Up to 22,528 lanes PCIe Gen 6 total; CXL 3.1
NIC	NVIDIA® BlueField®-4 DPU CX9 Any compatible PCIe NIC	BlueField-4 DPUs CX9 Any compatible PCIe NIC
Confidential Computing	Yes	Yes
Form Factor and Cooling	1S and 2S Server Air or liquid cooled 250 W to 450 W configurable	48U MGX rack 100% liquid cooled

1. Preliminary information. All values are up to and subject to change.
2. 200 TB recommended config.

Ready to Get Started?

To learn more about NVIDIA Vera, visit nvidia.com/en-us/data-center/vera-cpu/

NVIDIA