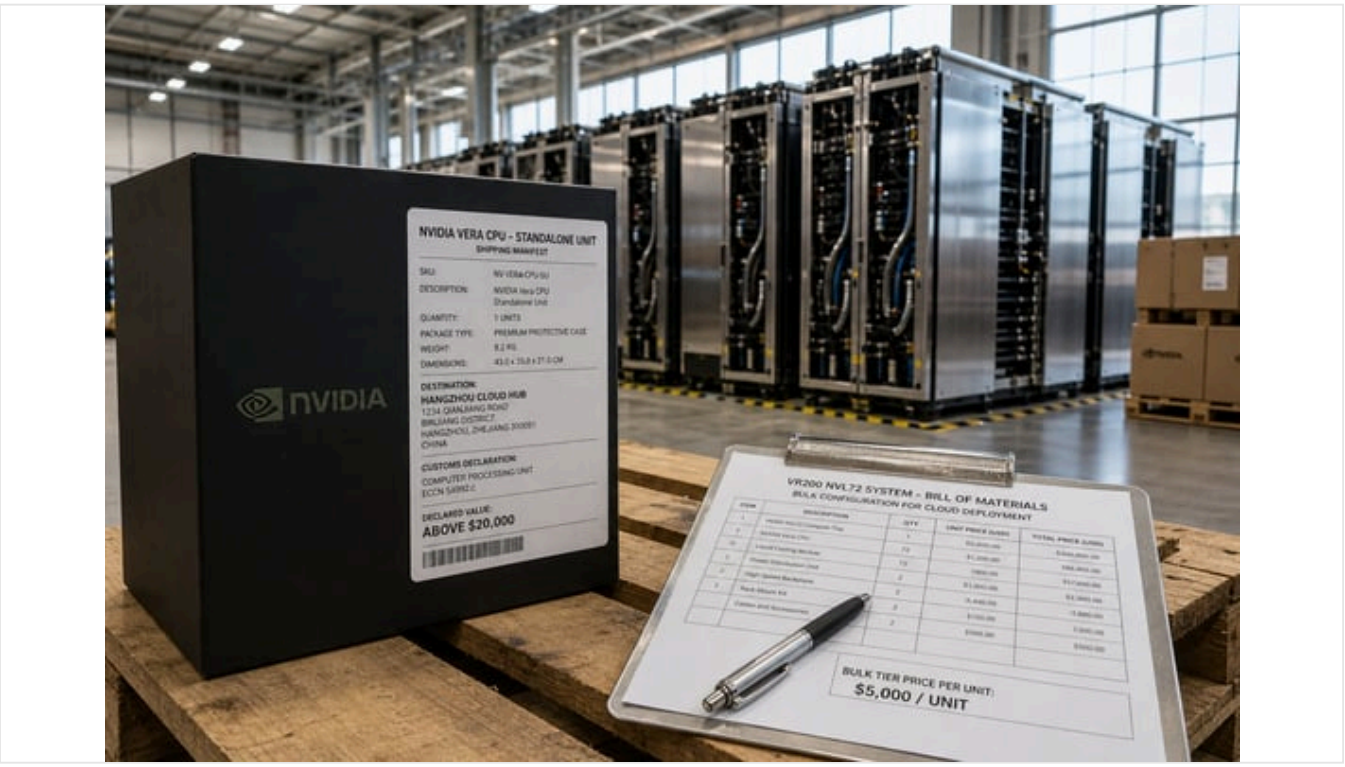
NVIDIA Vera CPU Price and Specs: What to Know

Published July 18, 2026 38 min read



Executive Summary

NVIDIA's **Vera CPU** is the company's first fully custom, in-house-designed data center processor, built around 88 **Olympus** cores that expose 176 hardware threads through NVIDIA's Spatial Multithreading scheme (Source: datacenterdynamics.com). It succeeds the Arm Neoverse V2-based Grace CPU and pairs with the Rubin GPU inside the **Vera Rubin** platform that NVIDIA CEO Jensen Huang put into "full production" at CES 2026 in Las Vegas (Source: blogs.nvidia.com). Vera delivers up to 1.2 terabytes per second (TB/s) of LPDDR5X memory bandwidth and up to 1.5 terabytes (TB) of memory capacity per socket, roughly triple Grace's 480 gigabytes (GB), while running on a single monolithic compute die that NVIDIA's second-generation Scalable Coherency Fabric (SCF) links to 3.4 TB/s of bisection bandwidth (Source: nvidia.com) (Source: servethehome.com).

On price, the picture is genuinely two-tiered and worth stating plainly because sources disagree by roughly 4x depending on volume and channel. Citing Morgan Stanley Research estimates, Tom's Hardware reports that NVIDIA plans to charge hyperscale customers around $5,000 per Vera CPU and $55,000 per Rubin GPU when the chips ship in volume inside VR200 NVL72 racks, pushing a full rack's bill of materials to roughly $7.8 million (Source: tomshardware.com). Outside that bundled hyperscaler channel, Reuters reports that a single standalone Vera processor "will cost 'well north' of $20,000 before bulk discounts, and a fully configured rack of 256 chips would run to around $10 million," according to SemiAnalysis, figures most relevant to smaller-volume buyers such as Chinese cloud firms now being pitched the chip as export-control-compliant compute (Source: reuters.com). NVIDIA's own leadership has framed Vera as opening a "$200 billion" total addressable market and is targeting close to $20 billion in combined Grace and Vera CPU revenue this fiscal year, which would require selling roughly 4 million Vera units and would make NVIDIA the world's largest CPU supplier by revenue, ahead of AMD and Intel combined (Source: tomshardware.com).

Independent benchmarking backs up at least some of the performance claims. Phoronix, given early access to pre-production hardware, measured Vera beating a single Intel Xeon 6980P "Granite Rapids" processor by 1.55x and an AMD EPYC 9575F by 10% on a geometric-mean basis across permitted workloads, while improving 1.63x generation-on-generation over Grace (Source: phoronix.com). AMD has since published its own modeled, not measured, rack-level comparisons claiming its upcoming 256-core "Venice" Zen 6 EPYC processor could beat Vera by 3.3x on a fixed power budget (Source: hothardware.com), a reminder that the competitive picture is still unsettled ahead of general availability. Commercial availability from OEM partners including Cisco, Dell, HPE, Lenovo, and Supermicro is expected in the second half of 2026, a timeline Phoronix independently

confirmed after early hands-on testing (Source: phoronix.com), and CoreWeave announced on June 1, 2026 that it had become the first cloud provider to bring up and validate a full Vera Rubin NVL72 rack in production (Source: coreweave.com). Meta has separately signed a multiyear partnership covering large-scale Grace deployment now and potential large-scale Vera deployment in 2027 (Source: nvidianews.nvidia.com). This report details Vera's full specification sheet, walks through every publicly reported price point and the discrepancies between them, compares Vera against Grace and against x86 rivals, and traces the roadmap through Rubin Ultra in 2027.

Introduction and Background

Anyone searching for "NVIDIA Vera CPU price and specs" as of July 2026 is looking for a chip that exists in an unusual state: architecturally disclosed in detail, independently benchmarked on pre-production silicon, already the subject of multiple named hyperscaler and cloud-provider commitments, yet without a single published, universal list price. NVIDIA first signaled the Vera Rubin platform's naming and general shape at Computex in Taipei in 2024, before formally introducing the six-chip Vera Rubin architecture and confirming the platform had entered full production during Jensen Huang's CES 2026 keynote at the Fontainebleau Las Vegas in January 2026 (Source: en.wikipedia.org (Source: blogs.nvidia.com). The Vera CPU itself received a second, more technically detailed unveiling at GTC 2026 in March, where NVIDIA disclosed the chiplet floorplan, cache hierarchy, and Olympus core microarchitecture for the first time (Source: servethehome.com).

Vera matters to buyers and analysts for a specific reason: it is NVIDIA's first data center CPU built entirely on an in-house core design rather than a licensed Arm Neoverse core. Grace, launched roughly four years earlier, used Arm's off-the-shelf Neoverse V2 core; Vera instead runs on Olympus, a CPU core NVIDIA designed from scratch, its first since the Denver core project nearly a decade earlier (Source: servethehome.com). That shift lets NVIDIA capture more of the CPU's value and differentiate against off-the-shelf Arm silicon used by cloud providers such as AWS Graviton and Google Axion (Source: phoronix.com). Building a custom core rather than continuing to license Arm's premade designs was as much a financial decision as a technical one: ServeTheHome's reporting notes that the total cost of licensing just the Arm instruction set architecture is cheaper than licensing one of Arm's premade CPU core designs, since royalties on a premade core are a larger factor, a structure custom silicon designers including Apple and Qualcomm already exploit and one NVIDIA now uses to differentiate its CPU hardware from other Neoverse-based competitors (Source: servethehome.com).

Vera is deployed in three principal configurations: as the host CPU inside a Vera Rubin superchip alongside two Rubin GPUs; as a standalone CPU in dense, liquid-cooled racks or conventional single- and dual-socket servers for reinforcement learning (RL), analytics, and general enterprise workloads; and paired with NVIDIA's ConnectX-9 networking silicon for storage and confidential-computing use cases (Source: nvidia.com)(Source: tomshardware.com). This report answers the two questions in the target query directly, price and specifications, and also addresses the secondary questions buyers ask alongside them: Vera's release timeline, how the Vera Rubin superchip is built, how Vera compares to Grace, where Vera sits on NVIDIA's broader data center CPU roadmap, core count and architecture detail, the shape of the Rubin Ultra platform arriving in 2027, and NVIDIA's emerging share of the data center CPU market as of mid-2026.

NVIDIA Vera CPU Architecture and Technical Specifications

Vera is built around 88 NVIDIA-designed Olympus cores, each fully compatible with the Armv9.2 instruction set architecture (ISA), delivering 176 hardware threads through a feature NVIDIA calls Spatial Multithreading, which partitions physical core resources between two threads rather than time-slicing them the way conventional simultaneous multithreading (SMT) does (Source: servethehome.com) (Source: newsletter.semianalysis.com). NVIDIA describes Olympus as "the first fully custom data center CPU core from NVIDIA," built for sustained high instructions-per-cycle (IPC) throughput on memory-intensive, control-flow-heavy code, the kind generated by agentic AI tool calls and sandboxed code execution rather than dense matrix math (Source: developer.nvidia.com). Each Olympus core uses a 10-wide instruction fetch and decode frontend paired with a neural branch predictor capable of evaluating two taken branches per clock cycle (Source: developer.nvidia.com). Independent supply-chain research firm SemiAnalysis, which produces bill-of-materials (BOM) costing on data center silicon, further reports that Olympus widens its floating point unit to six 128-bit-wide ports versus four on Arm's Neoverse V2, adding support for 8-bit floating point (FP8) operations under the Scalable Vector Extension 2 (SVE2) instruction set (Source: newsletter.semianalysis.com).

Physically, Vera is not a monolithic 88-core slab of silicon in the traditional sense. SemiAnalysis's die-level teardown describes a mesh layout of 91 physical core sites arranged in a 7x13 grid, with up to 88 harvested as active, a yield strategy common in large server dies, packaged across six total dielets using TSMC's CoWoS-R advanced packaging: one reticle-sized 3-nanometer (nm) compute die carrying the cores and NVLink-C2C logic, four LPDDR5 memory dielets, and one PCIe Gen6/Compute Express Link (CXL) 3.1 input/output (I/O) die (Source: newsletter.semianalysis.com) (Source: newsletter.semianalysis.com). NVIDIA's own marketing describes this as a single "monolithic compute die" with adjacent, disaggregated memory and I/O dielets, a design choice the company says avoids the non-uniform memory access (NUMA) partitioning found in competing multi-tile server chips from AMD and Intel, so every core sits an equal logical distance from every other core, cache slice, and memory controller (Source: servethehome.com).

The cache and memory subsystem is where Vera's generational jump over Grace is largest. Each core carries 2 megabytes (MB) of private L2 cache, double Grace's 1MB, feeding into a unified L3 cache of roughly 162 to 164MB depending on the source, up from 114MB on Grace (Source: servethehome.com). NVIDIA's second-generation Scalable Coherency Fabric (SCF) connects all 88 active cores to that shared cache and memory subsystem with 3.4 TB/s of bisection bandwidth, and NVIDIA states the design sustains over 90% of peak memory bandwidth even under full load, with each core provisioned for up to 14 GB/s of memory bandwidth, roughly three times the per-core rate of a traditional data center CPU (Source: developer.nvidia.com). Memory itself moves to second-generation LPDDR5X, delivered on removable Small Outline Compression Attached Memory Modules (SOCAMM), the industry's first attempt at bringing low-power, field-serviceable memory modules into the data center rather than soldering them to the board; SemiAnalysis specifies the design as eight 128-bit-wide SOCAMM modules of 192GB each, totaling up to 1.5TB of capacity at up to 1.2 TB/s of bandwidth (Source: newsletter.semianalysis.com).

Connectivity to Rubin GPUs runs over second-generation NVLink-C2C (chip-to-chip), rated at up to 1.8 TB/s of cache-coherent bandwidth, enough to let software address a Vera CPU's LPDDR5X and a Rubin GPU's High Bandwidth Memory 4 (HBM4) as a single unified memory pool, a capability useful for key-value (KV) cache offload during long-context inference (Source: datacenterdynamics.com) (Source: newsletter.semianalysis.com). Vera also supports PCIe Gen6 and CXL 3.1 for general I/O, and adds native confidential computing support, a feature entirely absent on Grace; Phoronix's early testing confirmed Vera already carries upstream Linux support for Arm's Confidential Compute Architecture (CCA), letting a Vera-based system attest to and isolate a workload's memory even from the host operator (Source: phoronix.com). On power, Phoronix's early hands-on testing at NVIDIA's Santa Clara headquarters recorded a peak 450-watt (W) socket thermal design power (TDP) for the CPU tested, with the accompanying LPDDR5X memory drawing roughly 50W or less, for a combined package in the same ballpark as top-end AMD EPYC Turin and Intel Xeon Granite Rapids CPUs, whose TDPs alone can reach 500W before memory is even counted (Source: phoronix.com).

Software support has moved unusually fast for a pre-launch chip. Phoronix confirmed that GCC 16.1 and LLVM Clang 21 already carry Olympus-specific compiler optimizations, upstreamed by NVIDIA well ahead of general availability, and that Linux kernel 7.1 and later include the driver support needed to run Vera under standard Arm64 server distributions such as Ubuntu and Fedora using Advanced Configuration and Power Interface (ACPI) rather than Device Trees (Source: phoronix.com). Testing itself has already spanned a broad set of workload categories: Phoronix's published test plan for its first Vera round covers code compilation, Stream memory-bandwidth benchmarks, SVT-AV1 video encoding, Python interpreter performance, OpenJDK Java workloads, Zstandard compression, and ClickHouse database throughput, chosen specifically to probe branchy, general-purpose code rather than narrow synthetic peaks (Source: phoronix.com). Wikipedia's source-cited tracking of the platform corroborates NVIDIA's own release messaging, noting the Rubin generation, and by extension its paired Vera CPU, "is scheduled for release in Q3 of 2026," consistent with the second-half-2026 OEM availability window NVIDIA itself has given (Source: en.wikipedia.org.

Table 1 below summarizes the generational jump from Grace to Vera using figures independently compiled by ServeTheHome and corroborated by NVIDIA's own developer documentation.

FEATURE	GRACE CPU	VERA CPU
CPU core architecture	Arm Neoverse V2 (licensed)	NVIDIA custom Olympus (Source: servethehome.com)
Core / thread count	72 cores / 72 threads	88 cores / 176 threads (Spatial Multithreading) (Source: servethehome.com)
L2 cache per core	1MB	2MB (Source: servethehome.com)
Unified L3 cache	114MB	~162MB (Source: servethehome.com)
Memory bandwidth	Up to 512GB/s	Up to 1.2TB/s (Source: servethehome.com)
Memory capacity	Up to 480GB LPDDR5X	Up to 1.5TB LPDDR5X (Source: servethehome.com)
SIMD width	4x 128-bit SVE2	6x 128-bit SVE2, adds FP8 (Source: servethehome.com)
NVLink-C2C bandwidth	Up to 900GB/s (first-gen)	Up to 1.8TB/s (second-gen) (Source: datacenterdynamics.com)
Confidential computing	Not supported	Supported (Source: phoronix.com)

Every metric in *Table 1* moves in Vera's favor, and NVIDIA's own materials round this up to a claimed 2x generational performance improvement, though as the Comparative Context section below details, that figure describes vendor-selected workloads rather than a universal multiplier (Source: phoronix.com). The practical takeaway for a buyer evaluating specs is that Vera is not an incremental Grace refresh: it swaps the licensed core for a custom one, roughly triples memory bandwidth and capacity per core, and adds hardware confidential computing, all changes aimed squarely at agentic AI workloads that spend more time on branchy, memory-bound orchestration code than on the dense linear algebra a GPU handles.

NVIDIA Vera CPU and Vera Rubin Superchip Pricing

Pricing for a chip NVIDIA has not yet shipped in volume is inherently an estimate, and the estimates in circulation as of July 2026 diverge sharply depending on the channel and volume tier being described, a discrepancy worth stating plainly rather than collapsing into a single number.

The most-cited figure comes from Morgan Stanley Research analysis of NVIDIA's VR200 NVL72 bill of materials, reported by Tom's Hardware from a leaked analyst chart: NVIDIA is projected to charge hyperscale customers approximately $5,000 per Vera CPU and $55,000 per Rubin GPU when the chips are sold in volume bundled inside a VR200 NVL72 rack (Source: tomshardware.com). That estimate is echoed almost verbatim in a separate Tom's Hardware analysis citing Mercury Research president Dean McCarron, who confirmed the $5,000-per-CPU figure as the basis for projecting NVIDIA would need to sell about 4 million Vera units to hit its stated CPU revenue target (Source: tomshardware.com). At that per-chip price, and with three dozen-plus Vera CPUs installed per NVL72 rack alongside its full complement of Rubin GPUs (detailed in the next section), the CPU line item alone contributes roughly $180,000 to a rack whose total bill of materials Morgan Stanley pegs at approximately $7.8 million, up from about $4 million for the prior-generation GB300 NVL72 (Source: tomshardware.com). Memory now drives an outsized share of that total: Tom's Hardware reports that memory content, spanning LPDDR5X, HBM4, and NAND flash, accounts for around 25% of a VR200 NVL72's total cost, roughly $2 million per rack, up 435% from the memory bill on a GB300 NVL72, because rack-level LPDDR5X capacity triples to 54TB from 17TB (Source: tomshardware.com) (Source: tomshardware.com).

That $5,000 figure, however, is specifically a bundled, hyperscale-volume, rack-integrated price, and it should not be read as Vera's standalone list price. Reuters reporting on NVIDIA's push to sell Vera CPUs into China as a workaround for GPU export restrictions confirms the standalone figure directly, citing SemiAnalysis for the "well north" of $20,000 per-chip estimate before any volume discounts (Source: reuters.com). The roughly 4x gap between $5,000 and "well north of $20,000" is not necessarily a contradiction: it likely reflects the difference between a fully bundled, deeply discounted hyperscaler NVL72 allocation negotiated at massive scale and a smaller-batch or standalone CPU-only purchase, the kind a Chinese cloud operator testing compatibility with a few hundred servers would actually pay. Buyers evaluating Vera pricing should treat $5,000 per chip as a lower bound reserved for the largest committed customers, and figures in the $15,000 to $25,000-plus range as more representative of smaller-volume or standalone procurement.

Neither figure is an official NVIDIA list price; NVIDIA does not publish per-unit Vera or Rubin pricing on its public product pages, consistent with its long-standing practice for data center silicon sold through OEM and hyperscaler channels rather than direct retail. What NVIDIA does disclose is the scale of the opportunity it expects Vera to unlock. On the company's Q1 fiscal year 2027 earnings call, an NVIDIA finance executive described Vera as opening "a brand-new $200 billion" total addressable market (TAM) the company had never previously addressed, and said NVIDIA had "visibility to nearly $20 billion in total CPU revenue this year," a figure NVIDIA later clarified spans Grace and Vera sold within superchips, NVL72 systems, and standalone CPU racks (Source: tomshardware.com) (Source: tomshardware.com). Reuters independently corroborates that revenue target, reporting that "Nvidia expects $20 billion in revenue from Vera chip sales by the end of this fiscal year to end-January" (Source: reuters.com).

Beyond the CPU itself, buyers pricing a full Vera Rubin deployment need to account for the Rubin GPU it ships alongside. NVIDIA's own developer documentation states the full Rubin die packs 336 billion transistors, up from 208 billion on Blackwell, delivering 50 petaFLOPS (PFLOPS) of 4-bit floating point (NVFP4) inference performance per GPU (Source: developer.nvidia.com). Two Rubin GPUs and one Vera CPU form a Vera Rubin superchip, and a full NVL72 compute tray, holding two superchips, delivers 200 PFLOPS of NVFP4 performance, 14.4 TB/s of NVLink 6 bandwidth, and 2TB of fast memory, according to NVIDIA's technical blog (Source: developer.nvidia.com). SemiAnalysis's own architectural framing situates that one-CPU-to-two-GPU head node ratio within a wider industry pattern, noting AMD's competing roadmap similarly pairs "1 Venice CPU to 4 MI455X GPUs per compute tray," a reminder that CPU-to-accelerator ratios, not just per-chip prices, shape total rack cost (Source: newsletter.semianalysis.com).

Rack-Scale and Hyperscale Deployment Economics

Vera is rarely sold or deployed as an isolated chip; NVIDIA's go-to-market for the platform is built around the rack as the unit of compute, which materially changes how buyers should think about cost per unit of useful work rather than cost per chip. The flagship configuration is the Vera Rubin NVL72 rack, which houses 72 Rubin GPUs and 36 Vera CPUs connected through a 260 terabytes per second (TBps) NVLink 6 fabric, a bandwidth figure independently corroborated in DatacenterDynamics' own platform coverage (Source: datacenterdynamics.com). NVIDIA states that, versus its

prior-generation GB200 NVL72 built on Grace and Blackwell, the Vera Rubin NVL72 trains large mixture-of-experts (MoE) models with one-fourth as many GPUs and cuts inference cost per million tokens by roughly 90% (Source: nvidia.com) (Source: nvidia.com); separately, DatacenterDynamics reported NVIDIA's own claim that the platform's rack-scale inference throughput reaches 3.3 times the performance of the prior GB300 NVL72 (Source: datacenterdynamics.com). Those figures are vendor-supplied and based on projected workloads (NVIDIA's own footnotes describe them as preliminary and subject to change), but they explain why hyperscalers are willing to pay a materially higher absolute price per rack than for the prior generation: the claim is lower cost per unit of AI output, not lower sticker price.

Early production deployment data is already testing NVIDIA's rack-economics claims in practice. CoreWeave, a specialized AI cloud provider, brought up and validated a full Vera Rubin NVL72 rack in June 2026, becoming the first cloud provider to do so; that milestone, and the substantial platform engineering CoreWeave built to operationalize it at production scale, is detailed in the case studies section below. The pace of that validation is notable set against SemiAnalysis's broader observation that 2026 marks a genuine inflection point for data center CPU demand, with the firm noting that "as we go through 2026, the demands on datacenter CPU and DRAM are only getting stronger," driven overwhelmingly by reinforcement learning and agentic workloads rather than the web-serving traffic that shaped prior CPU generations (Source: newsletter.semianalysis.com).

NVIDIA has also continued to extend the platform since its CES 2026 unveiling. The company added an optional seventh chip beyond the original six: a Groq 3 LPU inference-accelerator rack designed to pair with Vera Rubin NVL72 for the lowest-latency, largest-context agentic inference workloads, which NVIDIA states packs 256 LPUs with 128GB of SRAM, 40 petabytes per second (PB/s) of memory bandwidth, and 640 TB/s of scale-up bandwidth per rack (Source: nvidia.com), a sign of how quickly the platform's bill of materials, and its addressable price, keeps growing even after launch.

For enterprise and standalone-CPU buyers who do not need the full GPU superchip, NVIDIA offers Vera in three additional form factors: a dense, liquid-cooled Vera CPU rack for maximum sandbox and agentic throughput per rack unit; standard dual-socket servers, connected over the second-generation NVLink-C2C fabric; and single-socket servers for smaller deployments (Source: nvidia.com). NVIDIA claims a dense Vera CPU rack can host more than 22,500 concurrent agentic "sandbox" environments per rack, a metric aimed squarely at reinforcement learning and code-execution workloads rather than traditional web-serving throughput (Source: developer.nvidia.com).

Geography adds another wrinkle to deployment economics. Reuters reporting found that NVIDIA has told Chinese clients that Vera "could be available as soon as August and that they can begin placing orders," a route NVIDIA can pursue because United States export controls target high-end AI GPUs far more directly than general-purpose CPUs (Source: reuters.com). One major Chinese cloud provider reportedly plans to order more than 300 test servers with two Vera chips apiece before committing to a larger deployment, and named partners in the reporting include Alibaba and ByteDance, even as Beijing simultaneously pushes domestic chip self-reliance (Source: thenextweb.com) (Source: thenextweb.com). NVIDIA CEO Jensen Huang has publicly signaled that the company's Vera revenue forecast likely includes China, telling reporters in Taipei he "would think so" when asked directly (Source: thenextweb.com).

Comparative Context and Market Positioning

Positioned against its own predecessor, Vera's advantage on paper is unambiguous: more cores, double the threads, roughly triple the memory bandwidth and capacity, and a purpose-built core rather than a licensed one, as detailed in *Table 1* above. ServeTheHome's analysis situates that shift in competitive context: Grace, the company notes, "has become one of the most important Neoverse V2-based chips on the market," but Neoverse V2 remains a licensable design available to any silicon vendor, a differentiation gap Vera's custom Olympus core is specifically built to close (Source: servethehome.com). The harder comparison is against x86 incumbents AMD EPYC and Intel Xeon, where the picture depends heavily on whose benchmarks a buyer trusts.

Phoronix's early, NVIDIA-supervised but independently conducted benchmarking is the most credible third-party data point available as of July 2026. Testing a pre-production Vera system against a single AMD EPYC 9575F (a 5.0 gigahertz (GHz) high-frequency Zen 5 "Turin" part) and a single Intel Xeon 6980P ("Granite Rapids," Intel's current flagship with 128 cores), Phoronix found Vera delivered a geometric-mean performance advantage of 10% over the EPYC 9575F and 1.55x over the Xeon 6980P, alongside a 1.63x generational improvement over Grace (Source: phoronix.com). Reviewer Michael Larabel, whose benchmarking history with ARM64 server chips predates Vera by over a decade, called it "the most performant ARM Linux server processor I have ever tested," noting Vera clearly outperformed cloud-native Arm designs from Ampere Computing and the custom silicon inside Google Compute Engine and Microsoft Azure (Source: phoronix.com) (Source: phoronix.com). Importantly, Phoronix flagged that NVIDIA restricted which benchmarks it was permitted to run to workloads aligned with Vera's intended use cases, meaning the results should be read as a favorable but not necessarily comprehensive picture, and power-consumption and clock-frequency monitoring were explicitly disallowed during this first round of testing (Source: phoronix.com).

AMD has pushed back hard on the framing that Vera is competitive with its future silicon. In a blog post titled "Agentic AI Needs Rack-Scale CPU Performance, AMD EPYC Delivers It Today," AMD's own site states that "under the modeled 100 kW rack scenario, AMD EPYC 9965 delivers an estimated 2.37x the rack-level throughput of the NVIDIA Vera baseline and roughly 1.6x that of Intel Xeon 6980P," with next-generation "Venice" projected to extend the Vera comparison to 3.30x (Source: amd.com). AMD frames this explicitly as modeled, not measured, data: all configurations are "normalized to a modeled 100 kW rack built on 2P (two-processor) platforms," a methodology choice that reflects deployable service capacity rather than isolated peak processor behavior (Source: amd.com). On a per-core basis, AMD projected its 64-core Venice part would run 27% faster than an 88-core Vera, and even a 96-core Venice, with fewer cores than Vera, would still edge it out by 11% per core (Source: hothardware.com). The benchmarks AMD chose, including SPEC CPU, SPECjbb, NGINX, redis-benchmark, Memcached, and a MySQL transaction-processing test, are also weighted toward general data center and high-performance computing (HPC) workloads rather than the specific agentic-AI sandbox and reinforcement-learning tasks Vera is optimized for, a mismatch independent reviewers flagged directly (Source: hothardware.com). AMD's strongest non-modeled argument is architectural rather than a raw-throughput claim: its chips run native x86-64 code, the instruction set the overwhelming majority of existing server software targets, while Vera's Arm-based design carries real, if often small, software-compatibility costs for workloads not already ported to Arm64 (Source: hothardware.com).

NVIDIA's own marketing claims add a further wrinkle worth flagging as a genuine discrepancy rather than smoothing over: NVIDIA's developer blog states Vera delivers "up to 50% faster agentic sandbox performance" compared to unspecified "competitive platforms," while NVIDIA's consumer-facing Vera CPU product page separately states Vera is "up to 80 percent faster" than "traditional CPU infrastructure" and up to 1.8 times faster for code compilation, Python tool chains, and software analysis specifically (Source: developer.nvidia.com) (Source: nvidia.com) (Source: nvidia.com). The two figures are not necessarily inconsistent (they likely reflect different baseline comparisons and workload mixes measured at different times), but a buyer should not treat either as a single, universal "Vera is X% faster" number; the honest summary is that NVIDIA's own claimed advantage ranges from roughly 1.1x to 1.8x depending on workload and comparison baseline, broadly consistent with the independently measured 1.1x to 1.55x range Phoronix recorded against current-generation x86 silicon.

Early coverage of the platform also used differing rack nomenclature that is worth reconciling for readers encountering older material alongside this report. DatacenterDynamics' description of the platform around its initial unveiling referred to a rack capable of 3.6 exaflops of FP4 inference and 1.2 exaflops of FP8 training under the name "Vera Rubin NVL144," a naming convention that appears to count each of Rubin's compute dies separately (Source: datacenterdynamics.com). By general availability, NVIDIA and its launch partners had settled on "Vera Rubin NVL72" as the flagship rack name for the same underlying 260 TBps NVLink 6 fabric that outlet had earlier described under the NVL144 label, an important reconciliation for buyers comparing specs sourced from different points in the platform's roughly two-year public disclosure timeline (Source: datacenterdynamics.com).

Data Analysis and Evidence

NVIDIA's Q1 fiscal year 2027 results, covering the quarter ended April 26, 2026, give the clearest official financial anchor for Vera's commercial stakes. NVIDIA reported record total revenue of $81.6 billion, up 85% year over year, with record Data Center revenue of $75.2 billion, up 92% year over year, of which $60.4 billion came from Data Center compute and $14.8 billion from Data Center networking (Source: nvidianews.nvidia.com) (Source: nvidianews.nvidia.com). The same release specifically credits the quarter's momentum in part to the announced Vera Rubin platform, describing the Vera CPU as "the world's first processor purpose-built for agentic AI" (Source: nvidianews.nvidia.com).

Against that backdrop, NVIDIA's stated ambition for Vera is to become the world's largest CPU supplier by revenue within a single fiscal year of launch. Tom's Hardware, drawing on Mercury Research data and analyst commentary from Dean McCarron, lays out the arithmetic: the entire x86 server CPU market is worth roughly $30 billion annually, so NVIDIA's stated target of nearly $20 billion in combined Grace and Vera CPU revenue this fiscal year would represent close to two-thirds of that traditional market's total size, even though Vera itself sells against x86, not within it, as an Arm-based alternative (Source: tomshardware.com). For scale, Intel's data center and AI (DCAI) division generated $16.8 billion in 2025 and AMD's data center segment generated $16.635 billion the same year, though CPUs are only a portion of each figure, since both units also report accelerator and networking revenue (Source: tomshardware.com).

Mercury Research's broader x86 server tracking, as reported by Tom's Hardware, shows AMD reaching 46.2% of x86 server CPU revenue share in Q1 2026, versus Intel's 53.8%, with AMD's server unit share crossing 33% and AMD controlling 38.1% of total x86 CPU market value across all segments (Source: tomshardware.com) (Source: tomshardware.com). Those same Mercury Research figures put AMD EPYC's average selling price (ASP) at approximately $1,325 and Intel Xeon Scalable Processor (Xeon SP) ASP at approximately $1,125, with the two companies together shipping nearly 20 million EPYC and Xeon SP units for data center systems in 2025 (Source: tomshardware.com) (Source: tomshardware.com). Against those ASPs, Vera's $5,000-plus bundled price point (and its far higher $20,000-plus standalone price) is a striking multiple: three to sixteen times the typical x86 server CPU ASP, a reflection of how much of Vera's value NVIDIA is capturing through the tightly integrated superchip and rack rather than the CPU die alone.

McCarron told Tom's Hardware Premium he considers NVIDIA's 4 million-unit Vera shipment target achievable, noting NVIDIA is "already on track to deliver a number very near that for its GB300 and Rubin systems in FY2027," implying the CPU volume follows directly from committed GPU shipments, since Jensen Huang has stated "every two" Rubin GPUs sold ships with one attached Vera CPU (Source: tomshardware.com) (Source: tomshardware.com). NVIDIA reported $145 billion in total supply commitments, inventory, and prepayments at the end of Q1 FY2027, a figure Tom's Hardware ties directly to comments from NVIDIA's finance leadership about "remaining front-footed in securing sufficient supply" for the ramp (Source: tomshardware.com).

The broader CPU-for-AI-factories thesis behind Vera is corroborated by supply-chain analysts tracking the category independent of NVIDIA's own framing. SemiAnalysis's 2026 data center CPU landscape report states plainly that Grace-class CPUs have already become a limiting factor for some GPU deployments, noting that current AI workloads "are currently being slowed by the Grace CPUs in GB200 and GB300," the exact bottleneck Vera's higher core count, cache, and memory bandwidth are designed to remove (Source: newsletter.semianalysis.com). The same report frames 2026 broadly as an inflection point for data center CPUs precisely because reinforcement learning and agentic workloads, not traditional web serving, are now the primary demand driver, a shift SemiAnalysis says is pushing AWS and Microsoft toward "massive CPU buildouts of their own Graviton and Cobalt lines of CPUs as well as purchasing even more x86 general purpose servers" (Source: newsletter.semianalysis.com). SemiAnalysis's own cost-modeling practice underscores how closely the category is now being tracked: the firm states it maintains "detailed costing and breakdowns of AMD Turin, Venice, Intel Granite Rapids, Diamond Rapids, NVIDIA Grace, Vera and hyperscale ARM CPUs," the same competitive set referenced throughout this report (Source: newsletter.semianalysis.com).

Table 2 below consolidates the price and volume figures gathered across this report, since no single source publishes them together.

METRIC	REPORTED FIGURE	SOURCE AND CONTEXT
Vera CPU, bundled hyperscale volume price	~$5,000 per chip	Morgan Stanley Research estimate via Tom's Hardware, priced inside VR200 NVL72 racks (Source: tomshardware.com)
Rubin GPU, bundled hyperscale volume price	~$55,000 per chip	Same Morgan Stanley estimate (Source: tomshardware.com)
Vera CPU, standalone / smaller-volume price	"Well north" of $20,000 before discounts	SemiAnalysis estimate, reported by Reuters in China market-entry context (Source: reuters.com)
VR200 NVL72 full rack, total bill of materials	~$7.8 million	Morgan Stanley Research via Tom's Hardware (Source: tomshardware.com)
256-chip Vera rack (non-superchip config)	~$10 million	SemiAnalysis estimate via Reuters (Source: reuters.com)
Vera CPU total addressable market	$200 billion	NVIDIA finance leadership, Q1 FY2027 earnings call (Source: tomshardware.com)
Combined Grace + Vera revenue target, FY2027	~$20 billion	NVIDIA, via Tom's Hardware and Reuters (Source: reuters.com)
Projected Vera unit shipments, FY2027	~4 million units	Mercury Research analysis via Tom's Hardware (Source: tomshardware.com)
AMD EPYC average selling price, 2025 to 2026	~$1,325	Mercury Research via Tom's Hardware (Source: tomshardware.com)
Intel Xeon SP average selling price, 2025 to 2026	~$1,125	Mercury Research via Tom's Hardware (Source: tomshardware.com)

The spread in *Table 2* underscores that "Vera CPU price" does not resolve to a single number and depends heavily on who is buying, at what volume, and inside which package. What is consistent across every source is the direction: Vera commands a substantial premium over commodity x86 server CPUs, a premium NVIDIA is betting buyers will pay for the tight GPU-CPU coherence, memory bandwidth, and software stack integration the Vera Rubin platform provides as a system rather than as discrete parts.

Case Studies and Real-World Examples

Meta: A Multiyear, Multigenerational CPU and GPU Partnership

On February 17, 2026, NVIDIA and Meta Platforms announced a multiyear, multigenerational strategic partnership spanning on-premises, cloud, and AI infrastructure, enabling what NVIDIA described as large-scale deployment of NVIDIA CPUs alongside millions of NVIDIA Blackwell and Rubin GPUs (Source: nvidianews.nvidia.com). The deal's CPU component splits into two distinct phases. First, Meta and NVIDIA confirmed what the companies call the first large-scale, Grace-only CPU deployment in the industry, delivering what NVIDIA describes as significant performance-per-watt improvements across Meta's data centers running Arm-based Grace processors in production today (Source: nvidianews.nvidia.com). Second, and directly relevant to Vera's adoption timeline, the companies disclosed they are separately collaborating on Vera CPU deployment "with the potential for large-scale deployment in 2027," a full year or more after Vera's initial hyperscale shipments begin, indicating Meta is treating Vera as a distinct, later-stage adoption decision rather than an immediate Grace replacement (Source: nvidianews.nvidia.com). Meta CEO Mark Zuckerberg framed the broader partnership around the Vera Rubin platform specifically, saying Meta was "excited to expand our partnership with NVIDIA to build leading-edge clusters using their Vera Rubin platform to deliver personal superintelligence to everyone in the world" (Source: nvidianews.nvidia.com).

CoreWeave: First Cloud Provider to Validate Vera Rubin NVL72 in Production

CoreWeave's June 1, 2026 announcement is the clearest evidence to date that Vera Rubin has moved beyond lab demonstrations into operational cloud infrastructure. The company stated it had completed rigorous system-level validation of the entire NVL72 rack-scale architecture, becoming, in its own description, the first AI cloud provider to bring up and fully validate the platform (Source: coreweave.com). NVIDIA vice president Ian Buck corroborated the milestone directly, calling Vera Rubin "the most capable AI platform NVIDIA has ever built" and crediting CoreWeave's "full-stack, end-to-end approach" from cooling to orchestration (Source: coreweave.com). Quantitative trading firm Jane Street, an early customer cited in the announcement, said through its head of Quantitative Research, Craig Falls, that CoreWeave's infrastructure track record across NVIDIA Hopper and Blackwell gave the firm confidence to commit to Vera Rubin, adding the firm was "excited about the efficiency gains at rack scale translating into faster training runs and shorter iteration cycles for our researchers" (Source: coreweave.com). The buildout also demonstrates the platform's dependence on a broader hardware ecosystem beyond NVIDIA silicon: Dell Technologies chairman and CEO Michael Dell called the PowerEdge XE9812 server platform used in the deployment "a direct validation of what enterprise-grade hardware can do when it's paired with the right operational expertise" (Source: coreweave.com).

China Market Entry Through Compute Rather Than Accelerators

Vera's China rollout illustrates how the chip functions as more than a technical product; it is also a geopolitical and regulatory workaround. With NVIDIA's advanced AI GPUs still largely blocked from the Chinese market under United States export controls, Reuters reports that "Nvidia's market share in China has effectively fallen to zero, its CEO Jensen Huang said in October," hurt by export controls and Beijing's push for self-reliance in key technologies, prompting the company to pitch Vera CPUs directly to Chinese customers as a legal alternative since export licensing focuses far more heavily on high-end accelerators than general-purpose server CPUs (Source: reuters.com). Interest has been real but cautious: reporting describes a major Chinese cloud firm ordering more than 300 test servers, at two Vera chips each, specifically to validate software compatibility before committing to any larger rollout, and notes any near-term deployments will run in these customers' overseas data centers rather than domestically, given lingering uncertainty over migration costs off domestic chips and Beijing's own self-reliance push (Source: thenextweb.com). The stakes of that calculus are visible in NVIDIA's broader China predicament: with advanced GPUs still blocked outright, a grey market has emerged in which smuggled Blackwell-class B300 servers reportedly sell for around $1 million inside the country, a price signal illustrating how much unmet demand exists for NVIDIA compute of any kind in the market Vera is now targeting (Source: thenextweb.com). Reporting on the standalone chip's earliest customers also notes that "Anthropic and OpenAI were among its first users" of Vera outside China, a detail consistent with NVIDIA's own framing of reinforcement-learning post-training, not conventional web serving, as the workload Vera exists to accelerate (Source: thenextweb.com).

AMD's Public Rebuttal as a Market Signal

AMD's decision to publish a dedicated blog post and rack-level performance model specifically rebutting Vera, rather than simply promoting its own roadmap in isolation, is itself a data point about how seriously competitors are taking NVIDIA's CPU ambitions. AMD's post frames the stakes directly, arguing "the density positioned as future-looking is already being exceeded with standard infrastructure available now," a pointed response to Phoronix's Vera coverage that ran only weeks earlier (Source: amd.com), and HotHardware's own coverage notes NVIDIA had given "Linux blog Phoronix exclusive access" to the pre-production system under tightly managed conditions rather than opening it to unrestricted testing (Source: hothardware.com). Independent technology press covering the exchange noted AMD's own comparisons were "specifically curated as well," an important caveat given AMD's clear commercial incentive to characterize a not-yet-shipping competitor unfavorably (Source: hothardware.com). SemiAnalysis's independent roadmap tracking shows AMD extending its response beyond the Venice-versus-Vera comparisons: the firm plans a new 8-channel Venice SP8 platform succeeding the EPYC 8004 "Siena" line, bringing up to 128 dense Zen 6c cores to smaller-socket enterprise deployments, a segment SemiAnalysis expects AMD to gain further share in as Intel pulls back from its own comparable low-power server line (Source: newsletter.semianalysis.com).

Implications and Future Directions

Vera's arrival reshapes NVIDIA's position in the data center from GPU vendor to full-stack systems supplier, and the implications extend well beyond the CPU line item on a purchase order. By pairing a custom CPU core with its GPUs, NVLink switches, ConnectX-9 networking, and BlueField-4 data processing units (DPUs), NVIDIA is selling coherence and predictable performance at the rack level as the product, not any single chip in isolation, a strategy that raises switching costs for hyperscalers already standardized on the NVIDIA software stack while simultaneously inviting the antitrust and vendor-lock-in scrutiny that tends to follow vertically integrated platforms (Source: developer.nvidia.com).

The roadmap does not stop at Vera and Rubin. NVIDIA has already disclosed Rubin Ultra as the platform's successor, arriving in the second half of 2027 and roughly doubling Rubin's inference performance to 100 PFLOPS of NVFP4, achieved in effect by connecting two Rubin compute dies together into a denser package (Source: datacenterdynamics.com) (Source: en.wikipedia.org (Source: en.wikipedia.org. Rubin Ultra is expected to ship inside far denser, higher-power "Kyber" rack architectures, a signal that NVIDIA's roadmap cadence, an entirely new platform generation roughly every twelve to eighteen months, shows no sign of slowing, which itself pressures rivals to match an unusually aggressive release tempo Phoronix explicitly flagged as a key variable in whether Vera's early lead holds (Source: phoronix.com). Beyond Rubin Ultra, NVIDIA's roadmap points to a further platform generation, code-named Feynman after physicist Richard Feynman, expected in 2028 and continuing to pair next-generation GPU silicon with Vera-derived CPU designs (Source: en.wikipedia.org.

For competitors, Vera's entry accelerates an already-underway shift in how data center CPUs are evaluated. SemiAnalysis's broader 2026 CPU landscape research frames this year as a genuine inflection point for the category, driven by reinforcement learning and agentic workloads that place unprecedented demand on general-purpose compute even as GPUs dominate headlines, with AMD's Venice, Intel's Diamond Rapids, and multiple hyperscaler-custom Arm chips (AWS Graviton, Microsoft Cobalt, Google Axion) all converging on similar architectural bets around core count, memory bandwidth, and chiplet disaggregation at roughly the same time as Vera (Source: newsletter.semianalysis.com). That convergence suggests the "best CPU" question increasingly resolves less to raw core count and more to how tightly a given CPU integrates with the accelerators and networking fabric it ships alongside, an area where NVIDIA's vertical control over both halves of the Vera Rubin pairing gives it a structural advantage its rivals cannot fully replicate without their own competitive GPU or accelerator line. NVIDIA's own CPU track record gives that structural bet some credibility beyond marketing promises: ServeTheHome's analysis notes that, whatever its architectural limitations, Grace has been "wildly successful" by unit volume and hyperscaler goodwill, the base NVIDIA is now trying to extend with Vera's broader enterprise and cloud ambitions (Source: servethehome.com).

Buyers should also expect pricing to compress meaningfully once Vera exits its initial hyperscale-allocation phase and reaches broader OEM channel availability in the second half of 2026, following the well-established pattern in which early-allocation, capacity-constrained AI silicon commands a premium that narrows as supply catches up to demand, a dynamic already visible in the roughly fourfold gap between bundled hyperscaler pricing and smaller-volume standalone pricing documented earlier in this report.

Frequently Asked Questions (FAQs)

What is the price of the NVIDIA Vera CPU? There is no single official list price. Reported estimates range from around $5,000 per chip when bundled in volume inside VR200 NVL72 racks sold to hyperscalers, according to Morgan Stanley Research figures relayed by Tom's Hardware, to "well north" of $20,000 per chip for smaller-volume or standalone purchases, according to SemiAnalysis estimates reported by Reuters alongside NVIDIA's China sales push (Source: tomshardware.com) (Source: reuters.com).

What are the NVIDIA Vera CPU specs? 88 custom Olympus cores, 176 threads via Spatial Multithreading, Armv9.2 compatibility, up to 1.2 TB/s of LPDDR5X memory bandwidth, up to 1.5TB of memory capacity, 2MB of L2 cache per core, roughly 162MB of shared L3 cache, 3.4 TB/s of SCF bisection bandwidth, 1.8 TB/s of NVLink-C2C bandwidth to Rubin GPUs, and native confidential computing support (Source: datacenterdynamics.com) (Source: servethehome.com).

When does the NVIDIA Vera CPU release? NVIDIA announced the Vera Rubin platform in full production at CES 2026 in January, gave Vera a deeper architectural unveiling at GTC 2026 in March, and Phoronix reported the chip "remain[s] on track for shipping in the second half of the year," consistent with OEM availability from Cisco, Dell, HPE, Lenovo, and Supermicro that NVIDIA itself targets for the second half of 2026 (Source: blogs.nvidia.com) (Source: phoronix.com).

What is the Vera Rubin superchip? It pairs one Vera CPU with two Rubin GPUs over coherent NVLink-C2C, with two superchips forming an NVL72 compute tray delivering 200 PFLOPS of NVFP4 performance, 14.4 TB/s of NVLink 6 bandwidth, and 2TB of fast memory (Source: developer.nvidia.com).

Is the NVIDIA Vera CPU faster than AMD EPYC or Intel Xeon? Independent Phoronix testing found Vera roughly 10% faster than AMD's EPYC 9575F and 55% faster than Intel's Xeon 6980P on a geometric-mean basis across permitted workloads, though AMD's own modeled projections claim its unreleased Venice EPYC processor will reverse that gap by a wide margin once it ships (Source: phoronix.com) (Source: amd.com).

How does the Vera CPU compare to the Grace CPU? Vera moves from 72 licensed Arm Neoverse V2 cores to 88 custom Olympus cores, doubles thread count to 176, roughly triples memory bandwidth and capacity, adds confidential computing, and doubles NVLink-C2C bandwidth, as detailed in *Table 1* (Source: servethehome.com).

What is NVIDIA's data center CPU roadmap? Grace (2023) preceded Vera (2026), which will be followed by Rubin Ultra's CPU pairing in the second half of 2027 and a further Feynman-generation platform expected around 2028 (Source: en.wikipedia.org.

What is the Rubin Ultra platform? Rubin Ultra roughly doubles Rubin's inference performance to 100 PFLOPS of NVFP4, functionally by connecting two Rubin compute dies, and is expected in denser "Kyber" rack designs in the second half of 2027 (Source: en.wikipedia.org.

What is NVIDIA's data center CPU market share? NVIDIA does not yet report standalone CPU share, but the company is targeting close to $20 billion in combined Grace and Vera CPU revenue this fiscal year against an approximately $30 billion x86 server CPU market, which, if achieved, would make NVIDIA the leading CPU supplier by revenue ahead of AMD's 46.2% and Intel's 53.8% shares of the traditional x86 server segment (Source: tomshardware.com) (Source: tomshardware.com).

Conclusion

NVIDIA's Vera CPU pairs a genuinely novel architecture, its first fully custom CPU core in nearly a decade, with a pricing structure that resists a single clean answer. The specs are well documented and independently corroborated: 88 Olympus cores, 176 threads, up to 1.2 TB/s of memory bandwidth, up to 1.5TB of capacity, and a monolithic-compute-die design that roughly triples Grace's memory subsystem while adding hardware confidential computing. The price is where buyers need to read carefully rather than anchor on a single headline number: bundled hyperscale volume pricing around $5,000 per chip sits alongside standalone or smaller-volume pricing "well north" of $20,000, a gap of roughly four times that reflects channel and commitment level far more than any change in the underlying silicon.

Independent validation is arriving faster than is typical for pre-launch data center silicon, from Phoronix's early benchmarking showing real, if workload-specific, advantages over current AMD and Intel parts, to CoreWeave's production-validated NVL72 rack, to Meta's staged Grace-then-Vera deployment plan. AMD's aggressive, if explicitly modeled rather than measured, competitive response confirms the stakes NVIDIA's rivals see in the category. What remains genuinely open as of July 2026 is whether NVIDIA's roughly $20 billion CPU revenue target and 4-million-unit shipment goal survive contact with actual production ramp, supply allocation, and a rapidly evolving competitive field that includes AMD's Venice, Intel's Diamond Rapids, and a growing roster of hyperscaler-custom Arm CPUs. Buyers evaluating Vera today should treat every performance and pricing figure in this space, vendor-supplied or analyst-estimated, as provisional and revisit them once general OEM availability and independent, unrestricted benchmarking arrive later in 2026.

Tags: nvidia vera cpu, vera cpu price, vera cpu specs, nvidia vera rubin, vera rubin nvl72, grace cpu comparison, nvidia data center cpu, rubin ultra, olympus cores

---