

NVIDIA Vera CPU HPC Performance Brief

Version 1.1

Table of Contents

NVIDIA Vera CPU for HPC....4
Introduction....4
Vera Construction....5
NVIDIA Vera CPU HPC Performance....7
Memory BW - Stream....8
Numerical Weather Prediction - WRF Model....9
Computation Fluid Dynamics - OpenFoam....10
Structural Dynamics and Crash Simulation - OpenCourant (OpenRadioss)....11
Sparse Linear Algebra - HPCG....12
Genomics - BWA....13
LANL Agentic AI Case Study....14
TACC HPC Benchmarking Case Study....16
NVIDIA Vera CPU Rack....18
MGX Rack and Tray Configurations....19
Cluster Performance....22
Configurations....23

List of Figures

Figure 1. Liquid Cooled Vera Rack for HPC infrastructure..... 5
Figure 2. Vera delivers more than 3.4X Memory BW per watt..... 9
Figure 3. Vera delivers 2X higher performance/watt on Weather Forecasting model..... 10
Figure 4. Vera delivers 2.4X higher performance/watt on OpenFOAM model..... 11
Figure 5. Vera delivers leading performance and performance/watt Vs competition..... 12
Figure 6. Vera delivers greater than 2x performance and performance/watt on HPCG..... 13
Figure 7. Vera delivers leading Performance/Watt on the full BWA dataset..... 14
Figure 8. Vera delivers 4.4x higher URSA performance than x86..... 15
Figure 9. Vera delivers 2.2x on MuST, 5.5x on PSDNS, and 7.5x on MILC versus the TACC x86 baseline..... 17
Figure 10. NVIDIA Vera Liquid Cooled(LC) Modular Tray..... 19
Figure 11. Vera LC MGX tray offers 2x higher compute nodes vs competition..... 19
Figure 12. Vera LC MGX Rack Reference Architecture..... 21
Figure 13. NVIDIA Vera CPU rack deliver 2X higher achievable memory BW per watt..... 22
Figure 14. NVIDIA Vera delivers up to 2.6x higher performance on large clusters..... 23

NVIDIA Vera CPU for HPC

Introduction

High-performance computing (HPC) demands an optimal balance of compute performance, memory bandwidth, low latency, and energy efficiency to accelerate increasingly complex scientific simulations. NVIDIA established its CPU foundation in HPC with the Grace CPU Superchip, introducing a high-bandwidth, power-efficient LPDDR5X memory architecture and the NVIDIA Scalable Coherency Fabric (SCF). Grace demonstrated strong performance and energy efficiency across scientific workloads including weather forecasting, computational fluid dynamics, ocean modeling, and genomics, establishing NVIDIA CPUs as a compelling platform for HPC applications, in broad deployment.

NVIDIA Vera builds on this foundation with a new CPU architecture developed for the demanding requirements of agentic AI, reinforcement learning, data processing, and HPC applications. The workloads across these segments have a strong need for high single-thread performance, high memory bandwidth, low latency, and efficient scaling across cores. Vera advances the Grace architecture with NVIDIA's custom Olympus cores, next-generation coherent fabric, and power-efficient memory subsystem to address these requirements. This performance brief dives into how Vera enables and excels with key HPC workloads including real-world customer case studies and cluster scaling.

NVIDIA Vera scales into a high-density, multi-node HPC infrastructure designed to support large-scale scientific computing. Its modular architecture brings together compute, memory, and networking in a compact, rack-scale design, enabling efficient scaling from individual nodes to large HPC clusters. The figure below illustrates Vera's CPU-to-rack architecture, with further details covered in the NVIDIA Vera CPU Rack section.

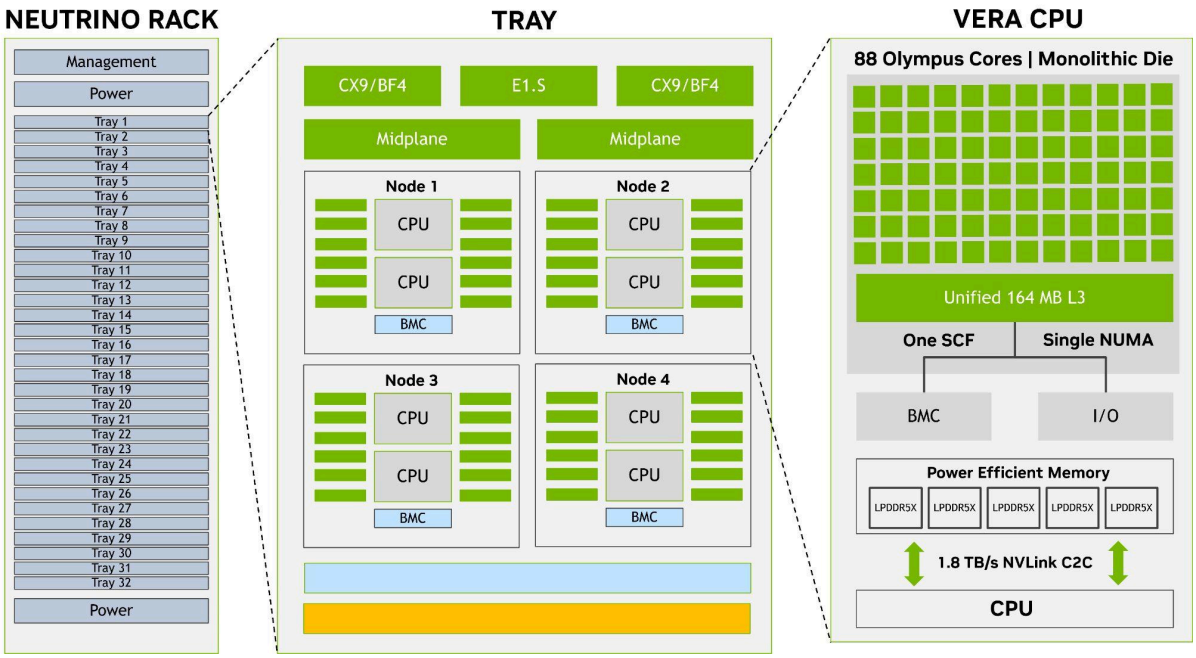


Figure 1. Liquid Cooled Vera Rack for HPC infrastructure

Vera Construction

Olympus Core

At the heart of NVIDIA Vera are 88 custom Olympus CPU cores, designed to deliver high instructions per cycle (IPC) and strong single-thread performance under full socket load. Olympus combines advanced branch prediction, 10-wide instruction decode, and deep out-of-order execution to keep more instructions in flight and minimize pipeline stalls. These capabilities are particularly valuable for HPC applications such as weather forecasting and computational fluid dynamics (CFD), which involve complex control flow, frequent data dependencies, and repeated execution of simulation timesteps. By maintaining high instruction throughput and reducing execution stalls, Olympus helps accelerate application execution and reduce overall time-to-solution.

Olympus also incorporates large private L1 and L2 caches and specialized hardware prefetchers to reduce memory access latency and keep execution resources productive. This is especially important for sparse linear solvers and unstructured CFD meshes, where indirect addressing, pointer chasing, and irregular data access patterns can leave CPU cores waiting for data. Combined with deep out-of-order execution, these features help sustain high per-core performance even under full socket utilization. For tightly coupled HPC workloads, consistent per-core performance reduces delays at synchronization points, improving parallel efficiency and overall simulation throughput.

NVIDIA Scalable Coherency Fabric and SOC

Vera integrates the second-generation NVIDIA Scalable Coherency Fabric (SCF), providing up to 3.4 TB/s of bisectional fabric bandwidth to connect Olympus cores, shared cache, memory controllers, and I/O resources. The fabric incorporates a large unified 164 MB system-level cache (L3) on a monolithic compute die, enabling low-latency data sharing and consistent latencies across cores. This architecture is especially important for HPC workloads such as OpenFOAM, and large-scale scientific simulations, where frequent synchronization, shared data access, and communication across processing cores can limit parallel scaling. By reducing inter-core latency and avoiding chiplet-related variability, Vera helps sustain predictable performance as HPC workloads scale. This predictability is critical for tightly coupled HPC workloads, where slower execution or delayed synchronization on even a few cores can stall the entire simulation, increasing overall time-to-solution.

SOCAMM2 LPDDR5X Memory and Energy Efficiency

The performance of HPC applications that process large datasets and frequently move data between memory and compute cores critically depend upon memory bandwidth and power efficiency. Vera uses SOCAMM2 LPDDR5X memory to deliver up to 1.2 TB/s of aggregate bandwidth and 14 GB/s per core across socket, providing substantial bandwidth for memory-intensive workloads such as weather modeling, CFD, and sparse matrix operations. Unlike traditional DDR5 and higher-power MRDIMM configurations, Vera's LPDDR5X architecture delivers high bandwidth with approximately 30–40 W of memory subsystem power, while maintaining enterprise-class ECC protection and modular serviceability. This combination delivers industry leading memory bandwidth per watt, reduces CPU-plus-memory platform power, and enables more compute capacity within the power and cooling constraints of large HPC deployments.

For a detailed overview of the Vera CPU architecture and its key innovations, refer to the NVIDIA Vera CPU Architecture Whitepaper.

NVIDIA Vera CPU HPC Performance

The following section details NVIDIA Vera CPU two-socket performance and power efficiency across key HPC workloads, spanning weather forecasting, computational fluid dynamics, scientific computing and customer case studies. These benchmarks highlight how Vera combines high-performance Olympus cores, high-bandwidth LPDDR5X memory, and a single NUMA low-latency coherent fabric to address demanding HPC workloads. The results showcase Vera's balanced architecture, delivering strong application performance and performance per watt for next-generation HPC systems.

NVIDIA Vera CPU is designed for HPC workloads that demand high memory bandwidth, low latency, and strong per-thread performance. For highly parallel, vector compute-intensive workloads, NVIDIA Blackwell and Rubin GPUs deliver massive computational throughput. Together, these architectures provide a complementary computing platform optimized for the diverse performance requirements of HPC applications.

Benchmark	**Vera Uplift Vs Latest Competition**
Stream - Memory BW Per Watt	3.4X
Numerical Weather Prediction- WRF	2.0X
Computation Fluid Dynamics - OpenFOAM	2.4X
OpenCourant (OpenRadioss)	1.4X
Linear Solver - HPCG	2.6X
Genomics - BWA	1.2X
TACC Case Study	7.5X

Memory BW - Stream

Memory bandwidth is a critical performance factor for many high-performance computing (HPC) workloads, where large datasets must be continuously moved between memory and CPU cores. STREAM is an industry-standard benchmark that measures sustained memory bandwidth through simple vector operations, including copy, scale, add, and triad. It provides a useful indication of how effectively a CPU platform can supply data to memory-intensive scientific applications such as computational fluid dynamics, weather forecasting, sparse linear solvers, and large-scale simulations.

NVIDIA Vera takes a different approach to server memory with power-efficient SOCAMM2 LPDDR5X, delivering up to 1.2 TB/s of memory bandwidth while consuming approximately 30–40 W for a fully populated memory subsystem. Traditional DDR5 systems offer lower aggregate bandwidth, while newer MRDIMM configurations can deliver higher bandwidth at the expense of substantially greater memory power and system cost. Fully populated MRDIMM systems can exceed 200 W of memory power, increasing the CPU-plus-memory platform power envelope and cooling requirements. Vera's LPDDR5X architecture provides a more balanced combination of high bandwidth, power efficiency, and memory capacity for HPC deployments.

This efficiency advantage is reflected in STREAM performance. As shown in the below figure, Vera delivers up to 3.4x higher sustained memory bandwidth per watt compared with AMD EPYC Turin, demonstrating how its LPDDR5X memory architecture translates high bandwidth into better power efficiency. For HPC clusters operating under fixed rack and datacenter power budgets, this enables more useful memory throughput within the available power envelope.

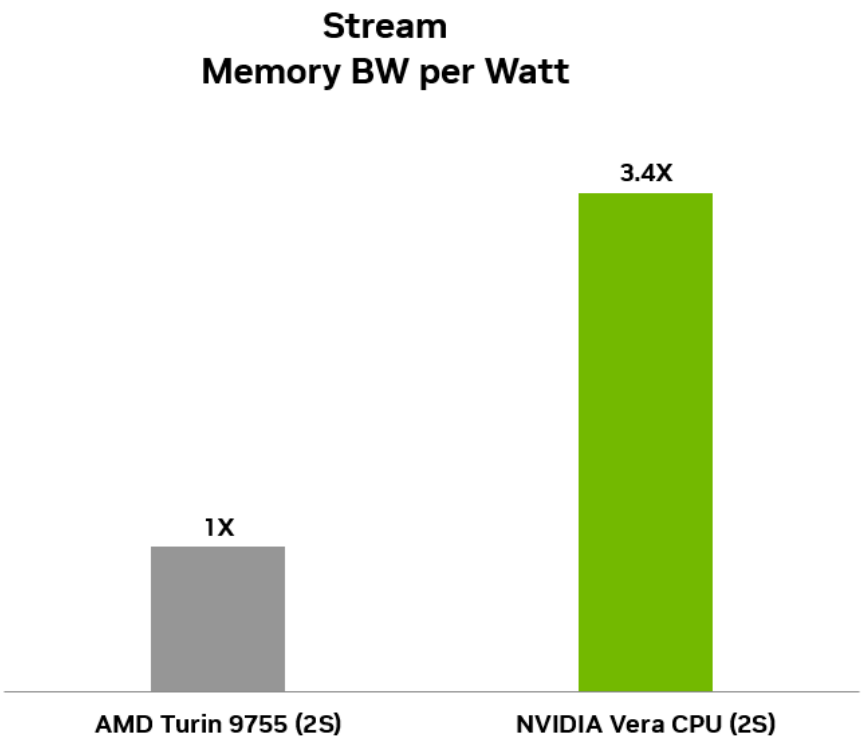


Figure 2. Vera delivers more than 3.4X Memory BW per watt

Numerical Weather Prediction - WRF Model

The Weather Research and Forecasting (WRF) model is a widely used numerical weather prediction system for atmospheric research and operational forecasting. WRF simulates atmospheric conditions by solving complex fluid dynamics and thermodynamic equations across large three-dimensional grids. These simulations involve intensive numerical computation, frequent memory accesses, and communication between processing cores, making WRF an important HPC workload for evaluating CPU performance, memory bandwidth, and time-to-solution.

WRF performance depends on how quickly the CPU can advance each simulation timestep across large atmospheric grids. Vera's high-IPC Olympus cores accelerate compute-intensive atmospheric dynamics and physics calculations, while high-bandwidth LPDDR5X memory sustains the frequent grid updates and data movement between timesteps. The low-latency coherent fabric further supports efficient data sharing across cores, helping reduce simulation time while maintaining power efficiency. Vera delivers 1.6X higher WRF performance and 2.0x better performance per watt compared with latest competition as shown below.

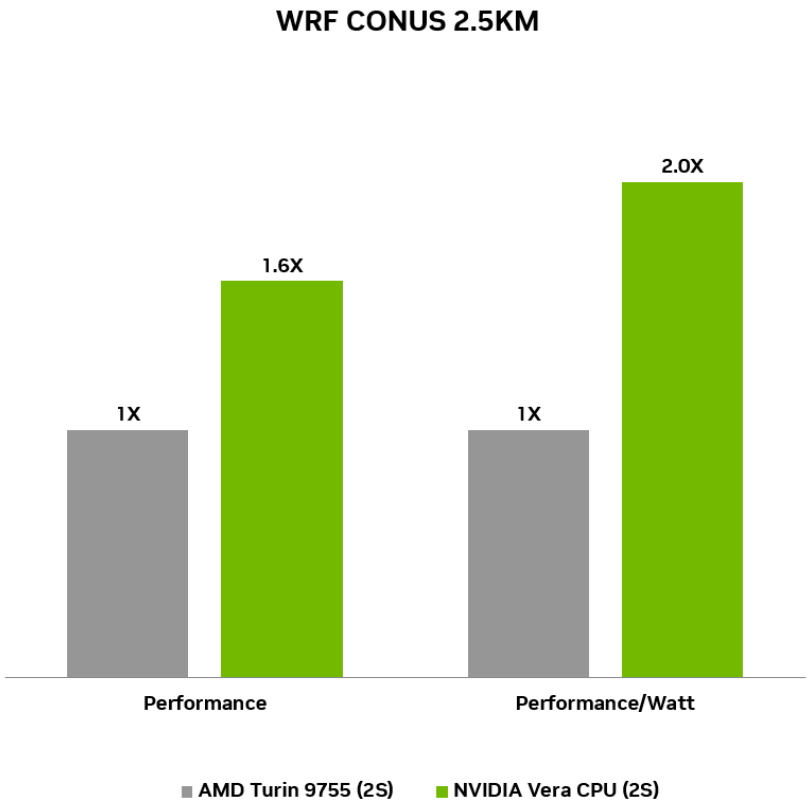


Figure 3. Vera delivers 2X higher performance/watt on Weather Forecasting model

Computation Fluid Dynamics - OpenFoam

OpenFOAM is a widely used open-source computational fluid dynamics (CFD) framework for simulating fluid flow, heat transfer, turbulence, and other complex physical phenomena. The motorBike Large dataset models airflow around a motorcycle using a large, complex computational mesh. The simulation involves iterative sparse linear solvers, frequent data movement, and communication across processing cores, making it a demanding HPC workload that stresses memory bandwidth, cache efficiency, and parallel scaling.

OpenFOAM performance is particularly sensitive to memory locality and NUMA topology as solver threads and MPI ranks access large, partitioned mesh data structures. Vera's monolithic compute die, unified 164 MB L3 cache, and single NUMA domain per socket simplify memory placement and reduce topology-related latency variability. Combined with high-bandwidth LPDDR5X memory, this architecture helps sustain efficient solver execution and scaling across cores. On the motorBike Large workload, Vera delivers 1.9x higher performance and 2.4x better performance per watt compared with latest competition.

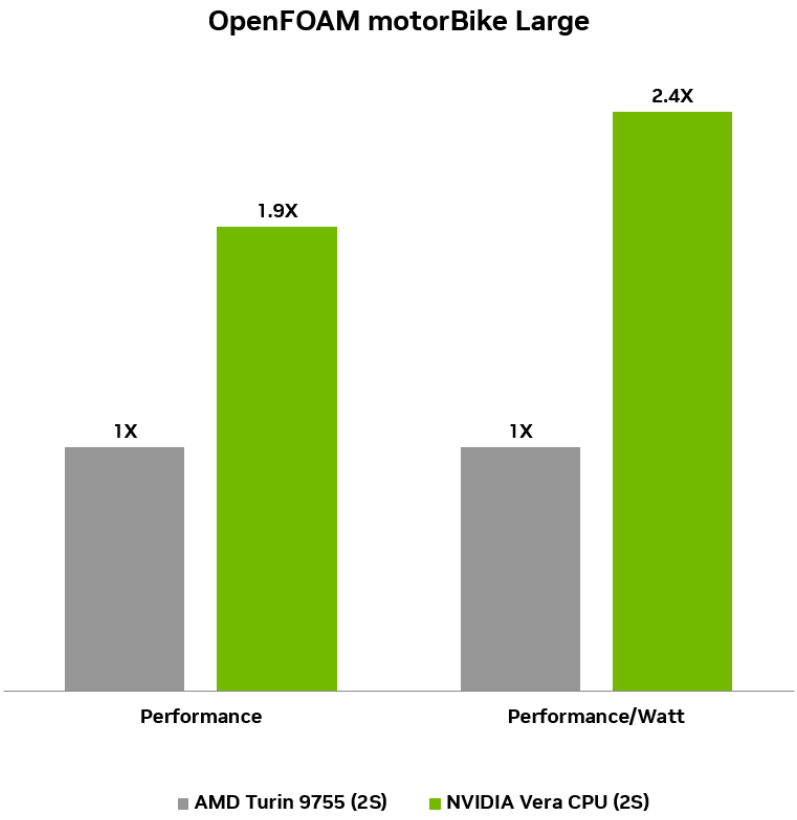


Figure 4. Vera delivers 2.4X higher performance/watt on OpenFOAM model

Structural Dynamics and Crash Simulation - OpenCourant (OpenRadioss)

OpenCourant is an open-source fork of OpenRadioss, created before the original project transitioned away from an open-source model. Built from the publicly available OpenRadioss codebase, OpenCourant preserves the underlying solver architecture and serves as a representative proxy for evaluating OpenRadioss-class HPC workloads.

OpenCourant is used for highly nonlinear structural dynamics simulations, including automotive crashworthiness, impact analysis, material deformation, and structural failure. These simulations use explicit time integration to calculate how millions of finite elements respond to forces and contact interactions over thousands of small timesteps. Each timestep involves intensive floating-point computation, element and contact processing, irregular memory accesses, and communication across MPI ranks and threads, making OpenCourant a demanding HPC workload for evaluating CPU compute performance, memory efficiency, and parallel scaling.

NVIDIA Vera is well suited for these computationally intensive simulations. Its high-IPC Olympus cores and wide execution engine accelerate element-level calculations and complex material models, while high-bandwidth LPDDR5X memory helps sustain data movement across large finite element meshes. Vera's monolithic compute die and NVIDIA Scalable Coherency Fabric further enable efficient data sharing and lower communication latency across cores, helping reduce synchronization overhead during parallel execution. Together, these architectural capabilities help improve simulation throughput, time-to-solution, and performance per watt for large-scale structural dynamics workloads.

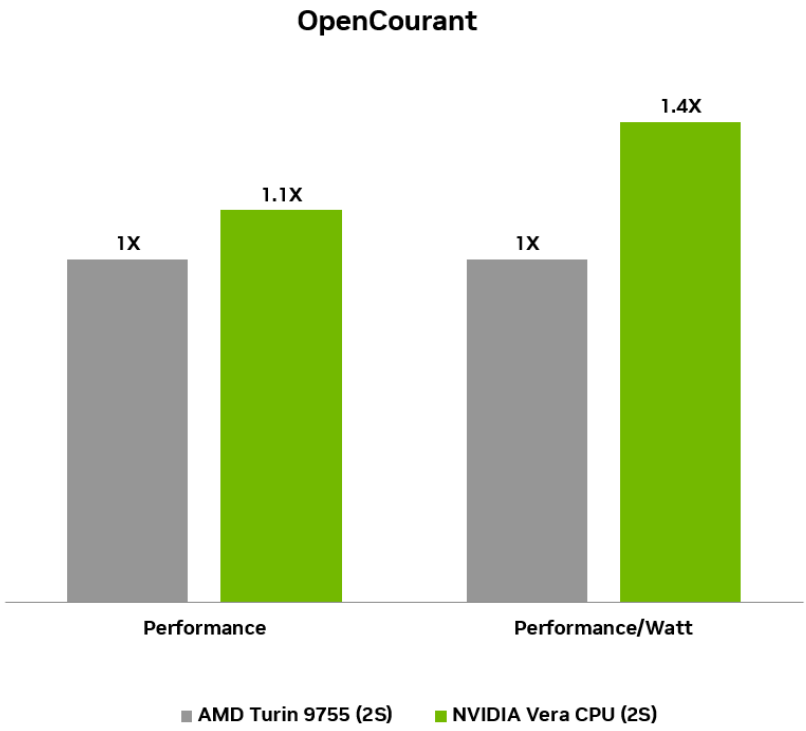


Figure 5. Vera delivers leading performance and performance/watt Vs competition

Sparse Linear Algebra - HPCG

HPCG (High Performance Conjugate Gradient) is a widely used HPC benchmark that represents the memory-intensive and communication-heavy behavior of scientific simulation workloads. It uses sparse matrix-vector operations, multigrid methods, and iterative solvers to evaluate how efficiently a computing system handles data movement, memory access, and inter-process communication. HPCG stresses memory bandwidth,

cache efficiency, communication latency, and parallel scaling, making it a valuable measure of real-world HPC system performance.

HPCG performance is particularly sensitive to memory bandwidth, NUMA locality, and communication overhead as processing cores exchange data during iterative solver operations. Vera addresses these challenges with up to 1.2 TB/s of LPDDR5X memory bandwidth, a high-bandwidth Scalable Coherency Fabric, and a unified 164 MB L3 cache. Its monolithic compute die and single NUMA domain per socket further reduce topology-related latency variability and simplify thread and memory placement, helping sustain performance across parallel solver operations. Vera delivers 2.1x higher HPCG performance and 2.6x better performance per watt compared with latest competition.

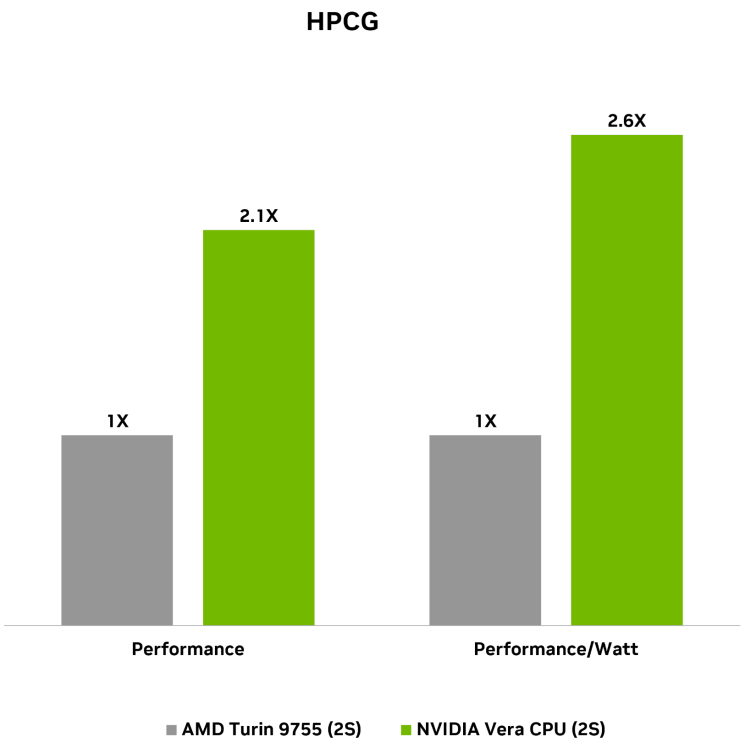


Figure 6. Vera delivers greater than 2x performance and performance/watt on HPCG

Genomics - BWA

BWA, the Burrows-Wheeler Aligner, maps DNA sequencing reads to a large reference genome, such as the human genome. It is a standard alignment step in whole-genome sequencing. Sequencing centers, clinical laboratories, and variant-discovery workflows such as GATK use BWA to produce the alignments that feed SNP and indel calling. This

brief runs the full BWA dataset. On that dataset the reads are independent, and each thread aligns its own reads without waiting on the others. BWA is an embarrassingly parallel workload: throughput follows the number of cores, rather than shared-memory latency or communication across cores.

The comparison is on dual socket Turin uses 256 cores while Vera uses 176, and the performance per core of Vera was 1.3x better than Turin. With independent reads, those extra Turin cores finish the alignment slightly sooner. Vera delivers higher efficiency at 1.2x the overall performance per watt.

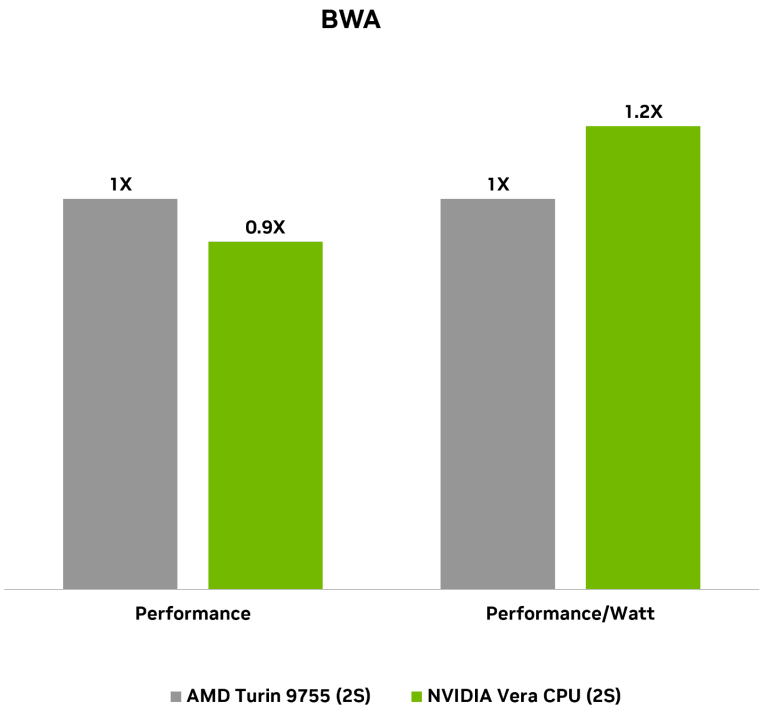


Figure 7. Vera delivers leading Performance/Watt on the full BWA dataset

LANL Agentic AI Case Study

URSA, the Universal Research and Scientific Agent, is an open-source scientific agent ecosystem created at Los Alamos National Laboratory. The paper URSA: The Universal Research and Scientific Agent describes a set of composable agents for planning, hypothesis generation, literature research, and computational execution. The execution agent writes files, runs commands, and iterates on the result, and it can call scientific simulation codes as tools. LANL shows the system planning and carrying out research

tasks, including a radiation-hydrodynamics design study, so that more of the work of setting up and running simulations sits with the agent.

In this end-to-end case study, the URSA workload is Parthenon Hydro, a compressible-hydrodynamics mini-app built on the block-structured adaptive mesh refinement framework developed at the laboratory. The task is to compile Parthenon Hydro and simulate a spherical blast wave in two-dimensions on both a uniform mesh and with adaptive mesh refinement. Both simulations start with mesh of 3.4 million elements and are advanced using Parthenon's RK-2 integrator. Compiling the mini-app and running both meshes stresses the host build, per-core numerical throughput, memory bandwidth for block data, and communication as refined blocks are created and exchanged. The LLM and inference infrastructure were the same on both Vera and on x86.

Performance is the wall time to compile Parthenon Hydro and run both blast configurations. Vera's high-IPC Olympus cores, high-bandwidth LPDDR5X memory, and low-latency coherent fabric address that mix of build, compute, and data movement. As shown in the below figure, Vera delivers 4.4x higher URSA performance than x86, reducing that time from 33 to 7 mins. Codex and Goose show the same direction of result, at 3.3x and 3.7x.

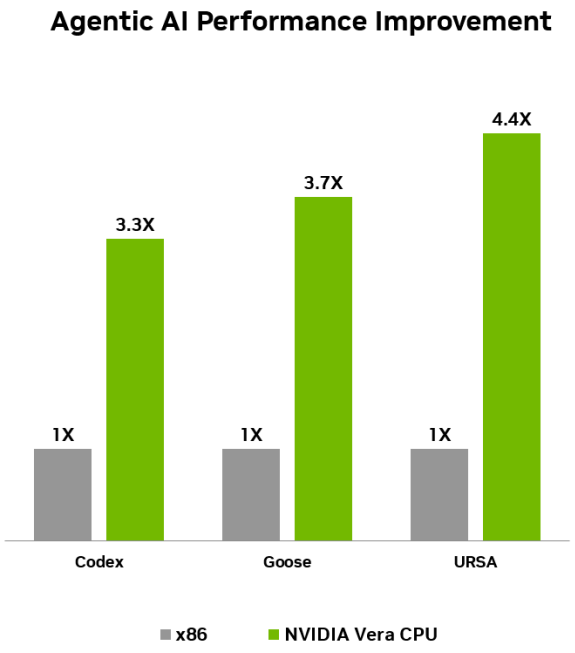


Figure 8. Vera delivers 4.4x higher URSA performance than x86

TACC HPC Benchmarking Case Study

TACC measured production applications on 2-socket, CPU-only nodes. Three of those codes are useful CPU benchmarks because each is a full application with a fixed problem, and each stresses a different part of the processor. MuST is a multiple-scattering code for the electronic structure of disordered materials. On the 56-atom CoCrFeMnNi alloy, most of the time is dense complex linear algebra, so the run measures floating-point throughput. MILC is a lattice QCD code. Its 48^3 x 64 benchmark is a standard procurement workload whose solvers are limited by memory bandwidth. PSDNS is a spatial direct numerical simulation of boundary-layer transition. At a 1408^3 grid, each timestep is a large structured-grid problem that stresses memory bandwidth and floating-point work.

As shown in below Figure, Vera delivers 2.2x higher performance than the TACC x86 baseline on MuST and 7.5x on MILC. On PSDNS, Vera is 5.5x that same baseline.

Vera is faster on all three because each code lands on a different strength of the processor, and the size of the gain follows that bottleneck. MuST spends most of its time in dense complex linear algebra on the 56-atom alloy. The Olympus cores sustain that floating-point work with high instructions per cycle, wide vector execution, and a 164 MB unified cache, which is what produces the 2.2x gain. MILC is limited by memory bandwidth. Its solvers stream the 48^3 x 64 lattice on every iteration, and SOCAMM2 LPDDR5X supplies up to 1.2 TB/s, so that data reaches the cores at the rate the solver needs. That bandwidth is the main reason for the 7.5x result. PSDNS combines the two. Each timestep on the 1408^3 grid is a memory-resident update: the cores do the floating-point work while the grid and the data shared across cores move through memory. LPDDR5X bandwidth feeds that update, and the Scalable Coherency Fabric on a single NUMA domain per socket keeps the sharing at low latency. That combination is what produces the 5.5x result.

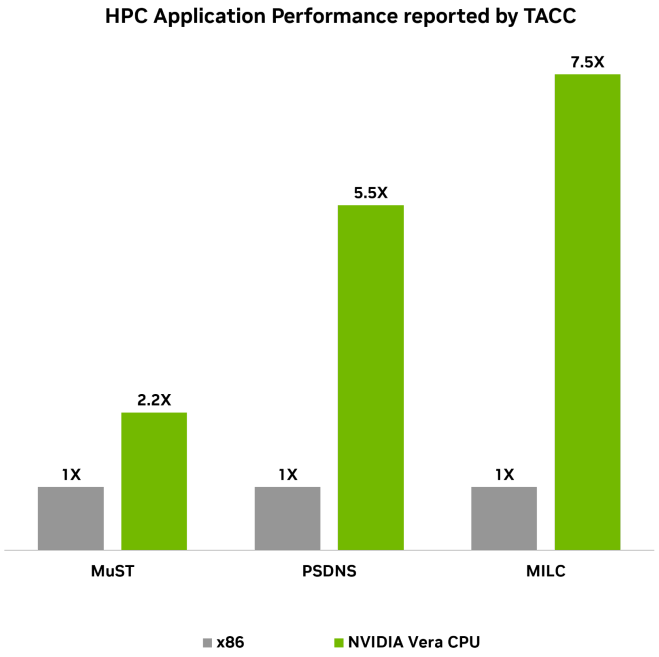


Figure 9. Vera delivers 2.2x on MuST, 5.5x on PSDNS, and 7.5x on MILC versus the TACC x86 baseline

NVIDIA Vera CPU Rack

Historically, high-performance computing (HPC) and commercial data processing infrastructure have evolved along separate paths. HPC systems relied on highly specialized architectures optimized for scientific computing, while enterprise and cloud datacenters prioritized standardized platforms, broad ecosystem support, and large-scale deployment. Today, these markets are converging as HPC, AI, and data processing workloads increasingly demand similar infrastructure capabilities: high compute density, massive memory bandwidth, high-speed networking, and energy-efficient liquid cooling.

The NVIDIA Vera HPC rack builds on this convergence, bringing the scale, standardization, and broad hardware ecosystem of modern datacenter infrastructure to scientific computing. Rather than relying on specialized, proprietary HPC systems, customers can deploy dense, liquid-cooled CPU infrastructure built on a common rack architecture supported by a multi-vendor ecosystem. This approach provides HPC customers with greater platform choice, supply chain flexibility, and a more scalable, cost-effective path to deploying next-generation supercomputing infrastructure.

The Vera liquid-cooled MGX rack is designed as a high-density CPU platform, with each 1U liquid-cooled compute tray includes four dual-socket Vera nodes, for a total of eight Vera CPUs per tray, with integrated ConnectX-9 for ethernet or InfiniBand networking, NVMe storage, and management plane. The trays are modular cableless designs shared across several MGX platforms enabling seamless integration, management and deployment. A full 32-tray rack scales to 256 Vera CPUs, enabling rapid deployment of large clusters without spreading capacity across many disaggregated systems.

MGX Rack and Tray Configurations

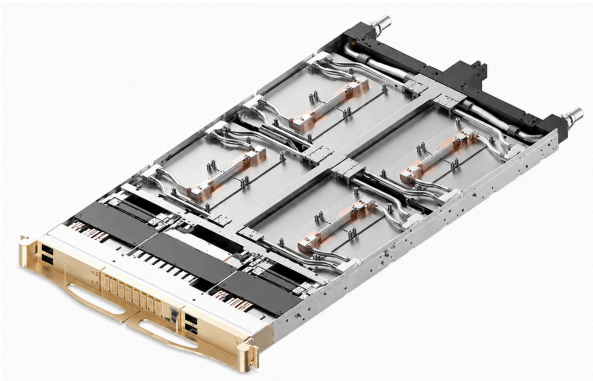
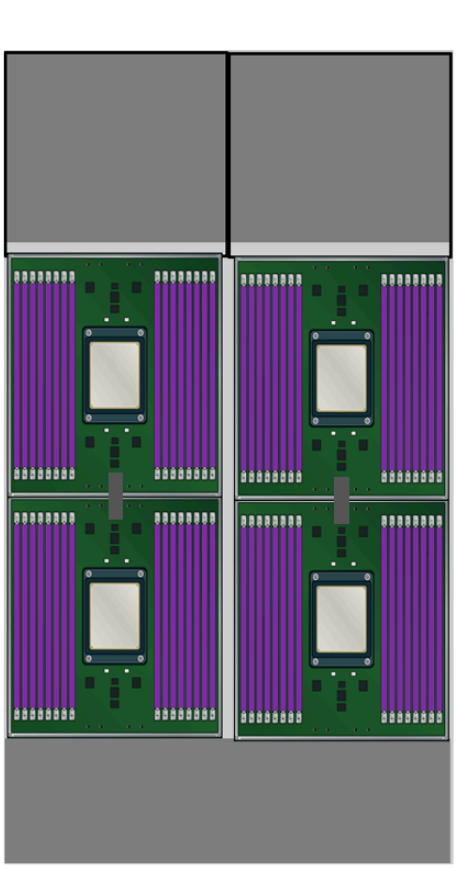
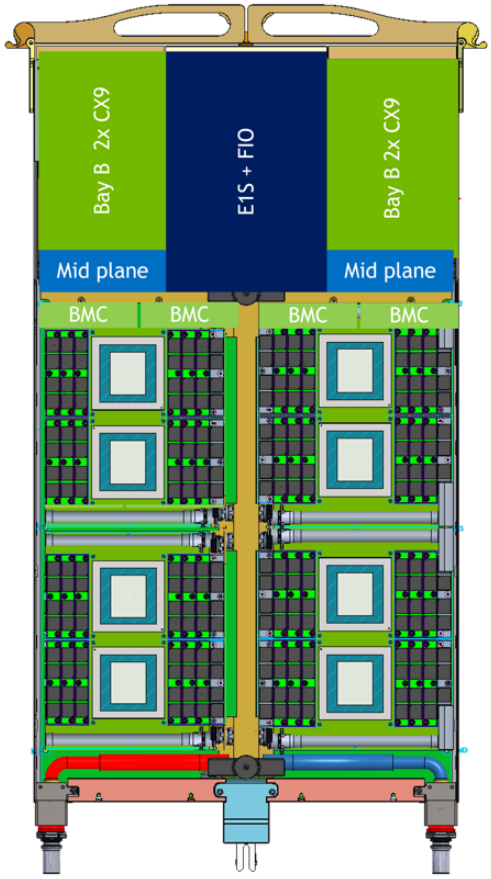


Figure 10. NVIDIA Vera Liquid Cooled(LC) Modular Tray



x86 Platforms
4 CPUs per Tray



NVIDIA CPU Platform
8 CPUs per Tray

Figure 11. Vera LC MGX tray offers 2x higher compute nodes vs competition

Vera Tray configuration

1U, 4-node Liquid-Cooled Compute Tray

- 8 CPUs per Tray
- 4x Vera 2S Nodes, each with 2x NVIDIA Vera CPUs
- 4x CX9/BF4 with 4x 800Gbps per Tray (N/S)

Per CPU

- Up to 1.5TB LPDDR5X Memory Capacity
- Up to 1.2 TB/s Memory Bandwidth
- 4 TB E1.S NVMe Disk
- 200 Gbit N/S Networking

LPDDR5X's compact, low-profile memory modules take up substantially less space than conventional DDR DIMMs and their sockets, leaving more room for compute. This smaller physical memory footprint helps fit eight NVIDIA Vera CPUs into a single liquid-cooled tray, compared with four CPUs in the DDR-based alternative shown here, doubling the number of CPUs per tray. SOCAMM2 LPDDR5X also reduces motherboard routing complexity and memory subsystem power, enabling a more compact and efficient server design.

The NVIDIA Vera CPU rack extends HPC computing to a high-density, fully liquid-cooled rack-scale architecture built on the NVIDIA MGX 1.2 platform. By integrating compute, power distribution, cooling, and management infrastructure into a common rack design, Vera enables dense CPU deployments while simplifying datacenter integration. The rack supports both 415–480 VAC and 800V DC power delivery options, providing flexibility to accommodate different datacenter power architectures. Additionally, other configurations including integrated switches in the rack will be supported.

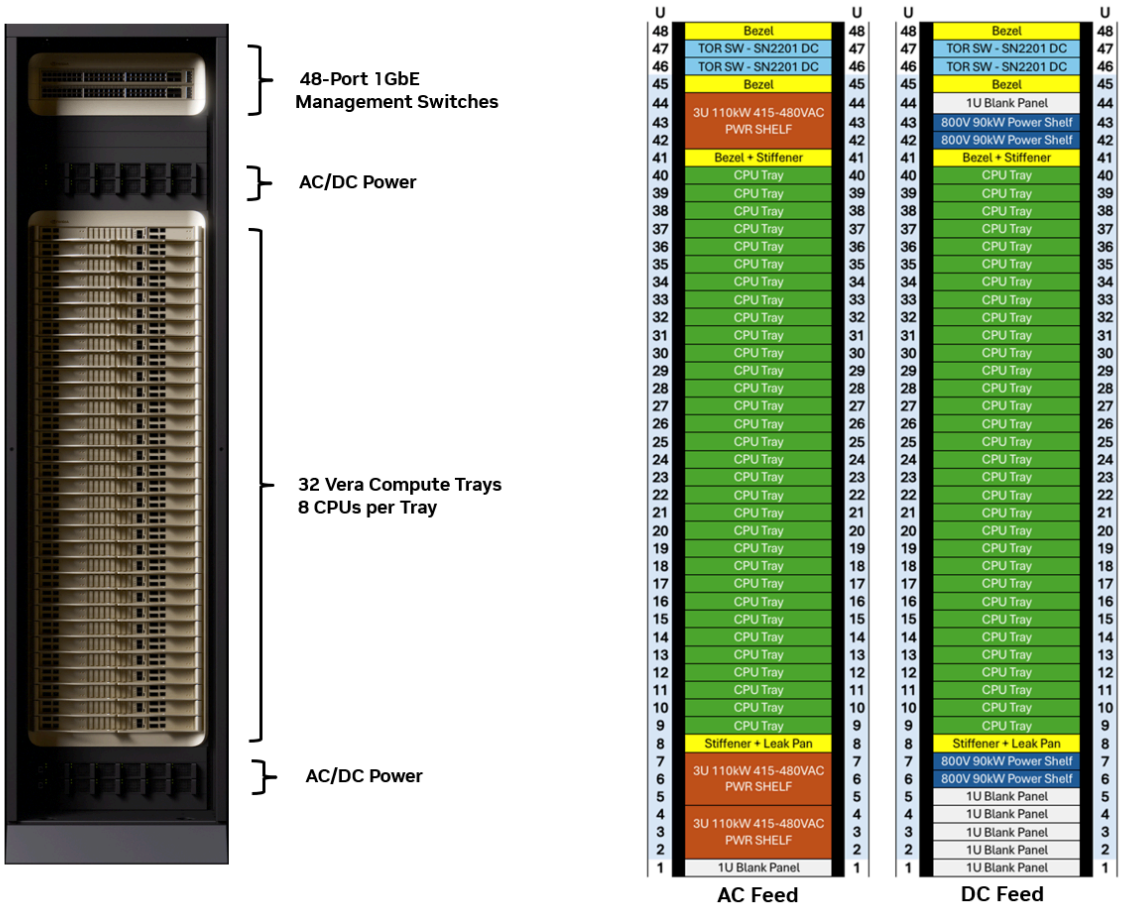


Figure 12. Vera LC MGX Rack Reference Architecture

The rack elevation optimizes space for compute trays while integrating power shelves, management switches, cooling manifolds, and mechanical infrastructure. Built for compatibility with NVIDIA MGX and NVL72 rack designs, Vera reuses common components including rack chassis, busbars, manifolds, and serviceable mechanical assemblies. For customers, flexible power delivery, liquid cooling, and MGX infrastructure reuse simplify HPC deployment and scaling while reducing total cost of ownership, with Vera offering 2X more nodes per rack compared to x86 based solutions.

As discussed in the previous section, memory bandwidth per watt is a critical metric for HPC workloads, particularly for memory-intensive applications operating within fixed datacenter power budgets. At multi-rack cluster scale, NVIDIA Vera's high-bandwidth, power-efficient LPDDR5X memory architecture delivers up to 2X higher achievable memory bandwidth per watt compared with next-generation x86-based racks. This advantage enables HPC customers to maximize aggregate memory throughput within

the same power envelope, improving infrastructure efficiency and scalability for large-scale scientific computing.

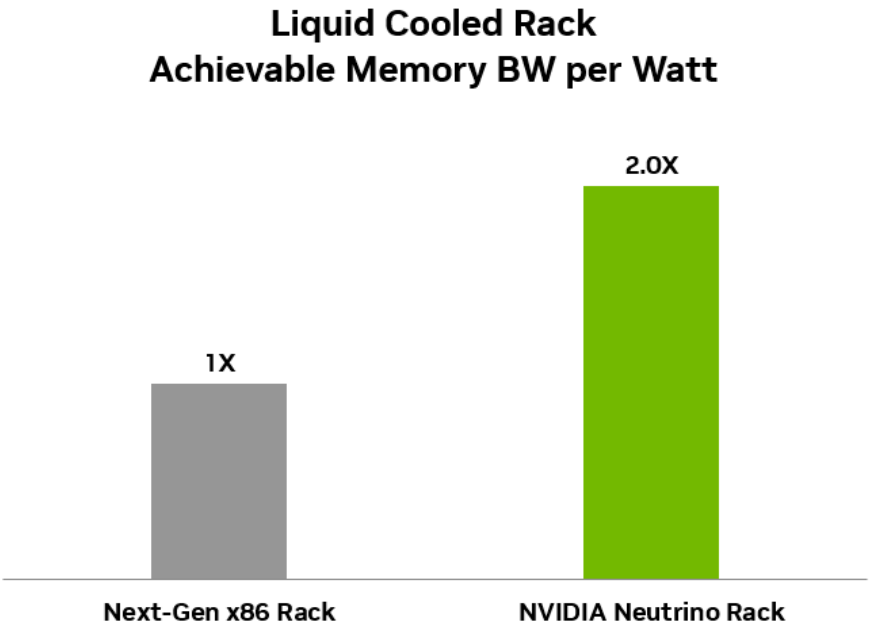


Figure 13. NVIDIA Vera CPU rack deliver 2X higher achievable memory BW per watt

Cluster Performance

Multi-node cluster performance is critical for HPC applications that distribute large-scale scientific simulations across multiple compute nodes. As workloads scale, overall performance depends not only on individual CPU capabilities but also on efficient communication, data exchange, and synchronization between nodes. NVIDIA Vera HPC clusters leverage high-speed NVIDIA InfiniBand networking to provide low-latency, high-bandwidth connectivity, helping minimize communication overhead and sustain parallel scaling for tightly coupled scientific workloads.

To evaluate multi-node scaling, three representative HPC workloads were executed on configurations of up to 12 Vera nodes. As shown in the figure below, NVIDIA Vera delivers up to 2.6X higher cluster performance compared with the prior-generation platform. Competitive clusters were not available for this evaluation. These results demonstrate Vera's ability to combine strong per-node performance with efficient inter-node communication, accelerating distributed HPC workloads and reducing overall time-to-solution.

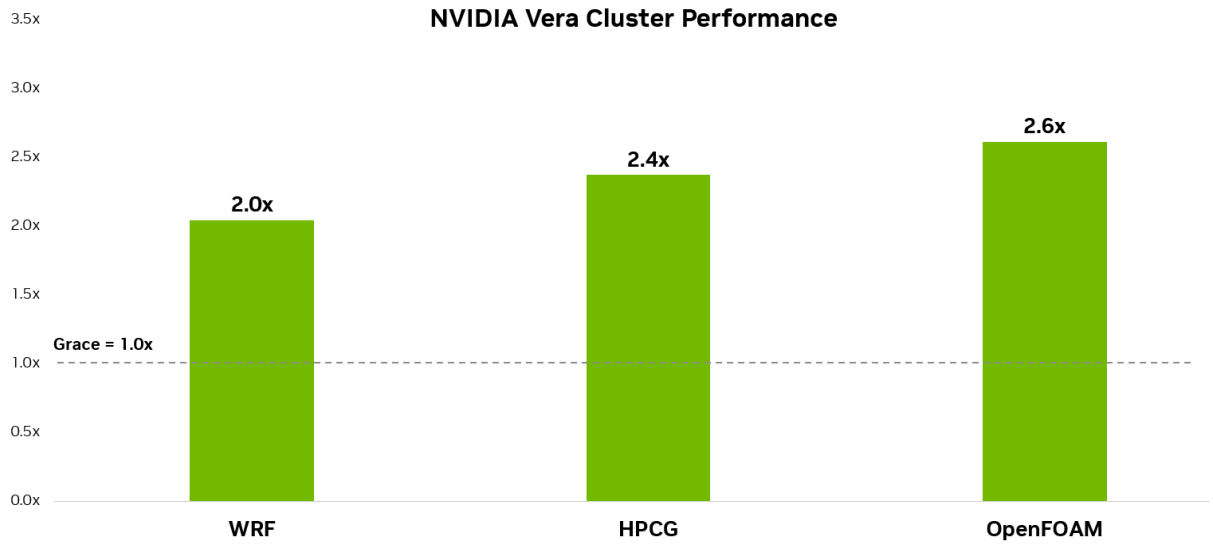


Figure 14. NVIDIA Vera delivers up to 2.6x higher performance on large clusters

Configurations

NVIDIA Vera Platform: C2 | Vera CPU | Socket Count: 2 | Cores / Threads: 88/176 | Memory: Samsung 1.5 TB 9600 MT/s LPDDR5X | Storage: KIOXIA KXD8DRJ93T84 3.5 TB | NIC: BlueField-3 | BIOS: NVIDIA NV_BIOS: 06.07.01.03 | OS: Ubuntu 26.04.1 LTS | Kernel: 7.0.0-2018-nvidia-bos-64k | Compiler: GCC16.2 | NVIDIA Grace Platform: C2 | Grace CPU | Socket Count: 2 | Cores / Threads: 72/72 | Memory: 480GB | Storage: Intel SSDPFUKX038T1 3.5TB | NIC: BlueField-3 | BIOS: American Megatrends International, LLC. Version 2.1a | OS: Ubuntu 24.04.4 LTS | Kernel:6.17.0-1026-nvidia-64k | Compiler: GCC16.2 | AMD Turin Platform: Supermicro H14DSG-OD | CPU: AMD EPYC 9755 | Socket Count: 2 | Cores / Threads: 128/256 | Memory: Micron 2.3 TB 6400 MT/s DDR5 | Storage: SAMSUNG MZQL23T8HCLS-00A07 3.5TB | NIC: E810-XXV | BIOS: American Megatrends International, LLC. Version 1.5 | OS: Ubuntu 24.04.3 LTS x86_64 | Kernel: Linux 6.8.0-83-generic | Compiler: GCC16.2 | Stream: https://www.cs.virginia.edu/stream/ | WRF:4.8.0 | OpenFOAM: v2606 | OpenCourant: latest-20261011| HPCG: https://developer.nvidia.com/nvidia-hpc-benchmarks-downloads | BWA: v0.7.19 | LANL URSA : v0.17.3 | TACC : MuST v1.9.5.4, PSDNS github.com/TACC, MILC v7.8.1

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2026 NVIDIA Corporation. All rights reserved.