Lenovo ThinkSystem SC777 V4 Neptune Server: Early Performance Overview

Planning / Implementation

From scientific simulation and engineering analysis to AI model training and inference, organizations increasingly require a common platform capable of supporting diverse workloads. The Lenovo ThinkSystem SC777 V4 brings together the technologies needed to accelerate both HPC and AI on a single system. The article will examine the platform design, discuss the types of HPC, AI, and AI-for-Science workloads it is designed to support, and present initial functional and application-level performance results.

The results presented in this paper provide an early look at performance on the SC777 V4 platform. Testing was conducted using pre-production laboratory systems and software environments available at the time of evaluation as the platform approaches general availability. As with any new platform, performance is expected to continue evolving as hardware, firmware, drivers, libraries, and applications mature. The results should therefore be viewed as directional indicators that highlight platform capabilities and workload behavior rather than final optimized performance levels.

ThinkSystem SC777 V4 overview

The Lenovo ThinkSystem SC777 V4 is designed from the ground up as a next-generation hybrid platform for both HPC and AI workloads—bringing together CPU, GPU, memory, and interconnect into a tightly integrated system.

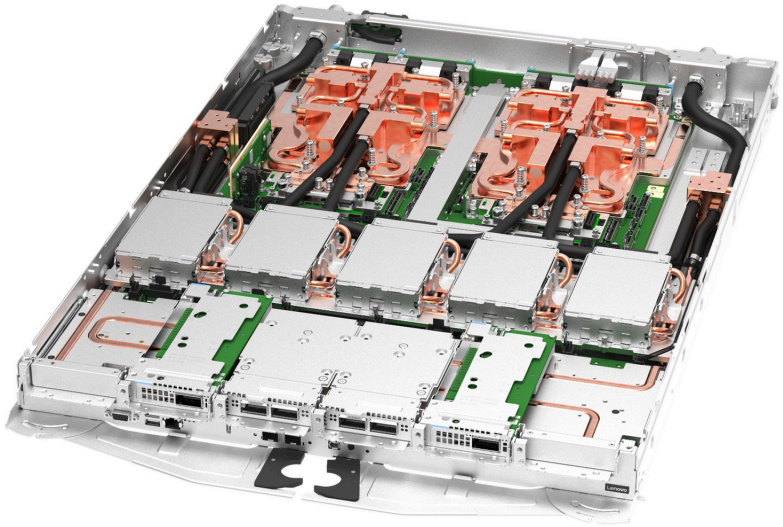

Figure 1. The Lenovo ThinkSystem SC777 V4 Neptune Server (AI Training configuration)

The SC777 V4 leverages the NVIDIA GB200 NVL4 architecture, combining Grace CPUs and Blackwell GPUs into a single, tightly integrated node design. Unlike traditional architectures that rely on discrete, loosely coupled components, SC777 V4 is engineered as a balanced, system-level platform.

The design of the SC777 V4 eliminates common bottlenecks seen in conventional systems:

- CPUs starved for memory bandwidth
- GPUs waiting on data movement
- Interconnect limiting multi-GPU scaling

As a result, the platform delivers:

- More consistent performance
- Higher system efficiency
- Better scaling across HPC and AI workloads

Each SC777 V4 server tray includes:

- **Compute**
  - 2x NVIDIA Grace CPUs (72 Arm Neoverse V2 cores)
  - 4x NVIDIA B200 GPUs (GB200-class modules)
- **Memory**
  - 960GB of LPDDR5x ECC Memory (480GB per CPU)
  - 744GB HBM3e Memory (186GB per GPU)
- **High-speed Interconnect**
  - 300 GB/s bi-directional NVLink-C2C connectivity between CPUs
  - 450 GB/s bi-directional NVLink-C2C connectivity between each CPU and its two integrated GPUs, unified coherent memory access
  - 600 GB/s bi-directional NVLink connectivity between each of the four GPUs
- **Storage & I/O**
  - Supports up to 10x E3.S SSDs (internal)
  - Support for NVIDIA NDR/XDR networking via CPU PCIe slots or GPU-direct modules, up to 800 Gbps per GPU with GPUDirect

Architecture

Features of the SC777 V4 that drive performance at scale include:

- High-bandwidth memory architecture feeds both CPU and GPU workloads efficiently
- CPU-to-GPU unified, coherent memory enables:
  - Faster data access
  - Reduced need for explicit data movement
  - Improved GPU utilization
- High-speed GPU-to-GPU interconnect:
  - Enables GPUs to operate as a single large compute domain
  - Critical for AI training and distributed HPC workloads
- Integrated into the ThinkSystem N1380 Neptune enclosure (13U)
- Designed for dense scaling:
  - Up to 96 GPUs per standard 19” width rack
- Flexible networking:
  - Supports high-speed NVIDIA NDR and XDR fabrics

The SC777 V4 compute nodes are installed vertically in the ThinkSystem N1380 enclosure, as shown in the following figure.

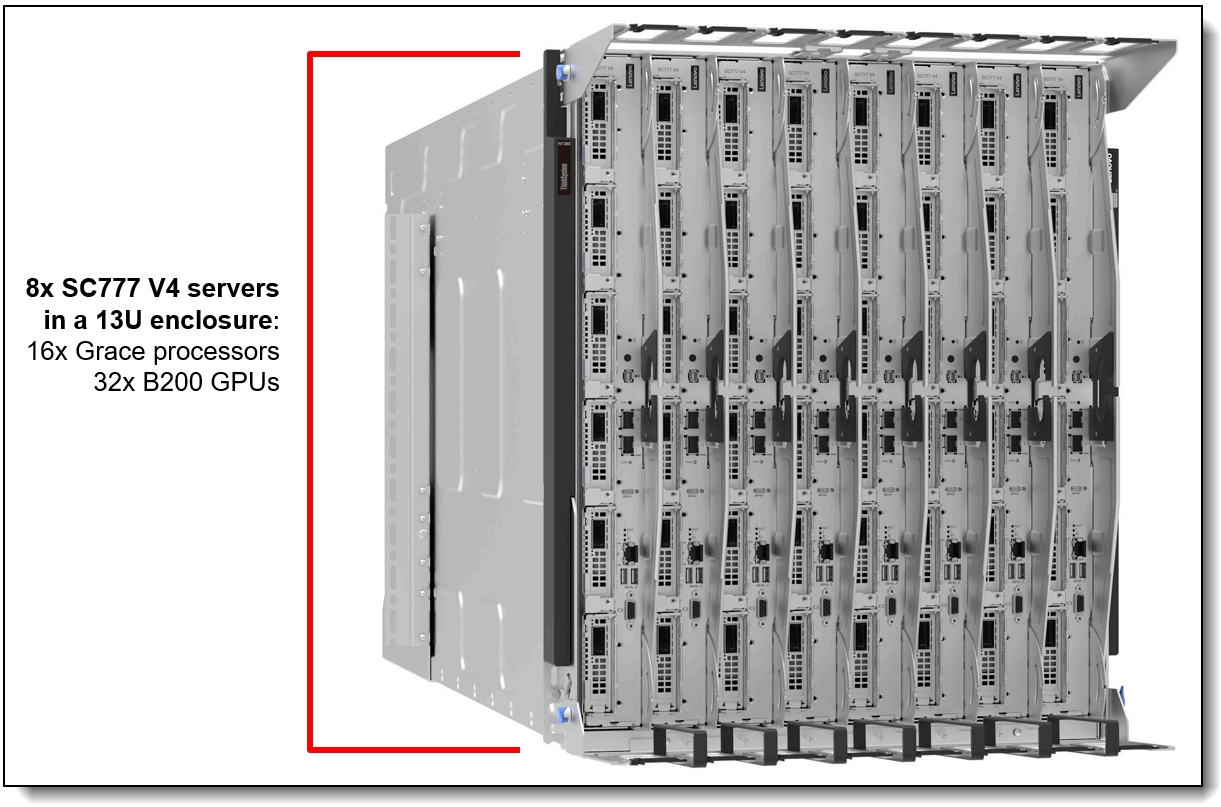

Figure 2. Front view of the N1380 enclosure with 8x SC777 V4 servers

The SC777 V4 incorporates Lenovo's Neptune liquid cooling technology, featuring a 100% direct water-cooled design that eliminates traditional airflow constraints and supports high power-density deployments.

- Enables higher sustained system performance under load
- Improves energy efficiency in dense GPU environments
- Supports inlet water temperatures up to 45°C, allowing some data centers to operate without chilled water systems, reducing overall energy consumption

SC777 V4 is not just a GPU-accelerated system - it is a purpose-built platform designed to unify HPC simulation and AI workloads on a single architecture.

Workloads

With its tightly integrated NVIDIA Grace + Blackwell architecture, the SC777 V4 is built to support a wide spectrum of modern workloads—spanning traditional HPC, enterprise AI, and emerging AI-driven scientific workflows.

- **Traditional HPC simulation and modeling**
  - Computational fluid dynamics (CFD)
  - Molecular dynamics
  - Electronic structure and quantum chemistry
  - Weather and climate modeling
  - Finite element analysis (FEA)
- **Artificial intelligence and data-driven workloads**
  - Large language model (LLM) training and fine-tuning
  - AI inference at scale
  - Computer vision and generative AI

    - Graph analytics
- **AI for Science and hybrid HPC + AI workflows**
    - AI-assisted simulation and surrogate modeling
    - Drug discovery and life sciences research
    - Materials science and quantum modeling
    - Climate prediction with machine learning augmentation
    - Real-time simulation and AI inference pipelines

Functional Baseline Performance

Functional performance tests are among the first validation steps performed after system installation and are used to baseline compute, memory, and interconnect behavior. These tests are not intended to represent real-world workloads. Instead, they provide a controlled way to verify that each component of the system is delivering expected performance and that the platform is properly optimized. These results also help inform expected behavior for real-world workloads when combined with application profiling data.

To establish these baselines, the following set of well-known synthetic benchmarks are typically used:

- Compute Foundational:
    - HPL (High Performance Linpack)
    - HPL-MxP (mixed precision)
    - GEMM / cuBLAS-based tests
- Memory Bandwidth Foundational:
    - STREAM
    - HPCG (compute + memory behavior)
- Interconnection Foundational:
    - NCCL (GPU-to-GPU communication)
    - NVBandwidth (CPU-to-GPU bandwidth)

The table below lists the functional performance baseline for the SC777 V4 platform.

The synthetic benchmark results are reported using three reference points:

- **Results** represent the measured performance achieved on the tested SC777 V4 configuration
- **Peak** represents the theoretical maximum performance obtainable from the underlying hardware based on architectural specifications
- **Expected** reflects the practical performance range anticipated for a well-configured and properly optimized system.

Comparing measured results against expected and peak values helps validate system health, identify potential performance bottlenecks, and confirm that the platform is operating as intended.

Table 1. Results of Synthetic Benchmarks

Tests	Results	Peak	Expected	Result vs Expected
CPU HPL (TF/s)	5.3	-	5.2	■■■■■■■■■■ 100%+
GPU HPL (TF/s)	149	160	~150	■■■■■■■■■□ 99%
GPU HPL with FP64 Emulation (TF/s)*	343	600	~345	■■■■■■■■■□ 99%
GPU HPL-MxP at FP16 (TF/s)	3954	-	-	-
CPU DGEMM (TF/s)	5.0	-	~5.2	■■■■■■■■■□ 96%
GPU GEMM at FP4 (TF/s)	6729	10000	-	-
GPU GEMM at FP8 (TF/s)	3001	5000	2907	■■■■■■■■■■ 100%+
GPU GEMM at FP64 (TF/s)	38.8	40	~40	■■■■■■■■■□ 97%

Tests	Results	Peak	Expected	Result vs Expected
CPU STREAM Triad (GB/s)	685	768	691	■■■■■■■■■□ 99%
GPU STREAM Triad (GB/s)	7508	8000	7360	■■■■■■■■■■ 100%+
CPU HPCG (GF/s)	114	-	-	-
GPU HPCG (GF/s)	4306	-	~4270	■■■■■■■■■■ 100%+
NCCL All-Reduce GPU-GPU (GB/s)	588	600	~600	■■■■■■■■■□ 98%
NVBandwidth CPU→GPU (GB/s)	212	225	201	■■■■■■■■■■ 100%+
NVBandwidth GPU→CPU (GB/s)	193	225	193	■■■■■■■■■■ 100%

* NVIDIA's FP64 emulation works by decomposing double-precision values into multiple lower-precision components, executing many high-throughput tensor core operations on those pieces, and then recombining the partial results to reconstruct an FP64 outcome. This leverages the massive performance gap between low-precision and native FP64 hardware, so the aggregate throughput of many fast low-precision operations can exceed native FP64 performance while aiming to preserve equivalent numerical accuracy.

Expected values are derived from publicly available data and internal performance projections informed by prior-generation results and published technical specifications.

Overall, the results show that the system is operating at or near (within 5%) expected performance levels across all major subsystems.

Real-World Workload Performance

While synthetic benchmarks are essential for validating system readiness, the true measure of a platform is how it performs on real-world workloads.

The results listed in the table below provide an early view of application-level performance across a representative set of HPC and AI workloads. Unlike synthetic benchmarks, these workloads exercise complex compute patterns, memory access behavior, and communication requirements that more closely reflect production environments. As such, they provide valuable insight into how the SC777 V4 performs across a diverse range of scientific computing and AI applications while highlighting the benefits of the NVIDIA GB200 NVL4 architecture.

The table presents an early view of application-level performance on the SC777 V4. Uplift represents the relative performance improvement compared to the baseline system, where 1.50x indicates a 50% improvement and 2.00x indicates twice the performance of the baseline.

Table 2. Real-World HPC and AI Workload Comparative Results

Workload	Results	Uplift Comparison (Up to)**
Gromacs benchPEP	13.81 ns/day	1.48x H100 80GB SXM | 1.41x GH200
Gromacs benchPEP-h	21.347 ns/day	1.59x H100 80GB SXM | 1.34x GH200
Gromacs benchRIB*	47.064 ns/day	1.47x H100 80GB SXM | 1.27x GH200
Gromacs benchMEM*	459.93 ns/day	1.76x H100 80GB SXM | 1.08x GH200
OpenMM Amber20 STMV	431.484 ns/day	1.76x H100 80GB SXM | 1.48x GH200
OpenMM Amber20 Cellulose	223.128 ns/day	1.45x H100 80GB SXM | 1.21x GH200
Quantum ESPRESSO AuSurf*	65.6 sec	1.47x GH200
MLPerf Training Llama2 70B Offline	19.77 minutes	3.2x H100 80GB SXM | 2.3x H200 141GB SXM
MLPerf Inference Llama2 70B Offline	49,382 tokens/s	2.8x H200 141GB SXM

* Measured on 1 GPU
** Comparisons are done with respect to a server with the same number of GPUs as the presented SC777

V4 result

While the SC777 V4 delivers strong performance improvements across both HPC and AI workloads, the magnitude of uplift differs based on workload characteristics.

The table below summarizes the key differences between traditional HPC and AI workloads, helping explain why some workloads realize larger gains from newer GPU architectures than others.

Table 3. HPC and AI Workload Characteristics

HPC Workload Characteristics	**AI Workload Characteristics**
FP64-dominant	Leverages lower precision compute
Memory-bandwidth and latency sensitive	Aligned with tensor core acceleration
Dependent on tightly coupled communication patterns	Benefit from higher GPU compute throughput and parallelism

Because HPC applications are often bound by memory access, data movement, communication patterns, and FP64 precision requirements, they generally realize more modest generational performance gains than AI workloads.

The SC777 V4 demonstrates strong, balanced performance across a diverse set of real-world workloads. HPC applications see steady, meaningful improvements driven by system balance and memory efficiency. AI workloads realize significantly larger gains, reflecting the rapid evolution of GPU architectures toward data-driven and mixed-precision computing.

Conclusion

The SC777 V4 delivers strong and consistent performance across both HPC and AI workloads. It provides meaningful gains for traditional HPC applications, while AI training and inference workloads benefit from significantly higher improvements. This reflects a broader industry trend, where modern GPU architectures are increasingly optimized for AI and mixed-precision compute. At the same time, the SC777 V4 maintains the performance, stability, and scalability required for demanding scientific workloads.

The SC777 V4 provides a single platform capable of supporting today's HPC workloads and enabling the next wave of AI-driven innovation.

The benchmark results presented in the document reflect a snapshot of performance under the tested configuration. Results may vary based on system settings, including operating system, BIOS/UEFI configuration, firmware levels, software environment, and workload conditions.

Authors

Kevin Dean is the Senior Manager of the HPC/AI Solution Architect Team in the ISG Offerings Group at Lenovo. Kevin oversees the strategy and processes for HPC and AI solutions in this position, contributes his knowledge of HPC/AI application performance, and serves as the CAE/manufacturing architect. Kevin has over 20 years' experience in HPC, AI, and computational engineering, including nine years at Lenovo focused on HPC/AI performance and solution architecture, and 12 years in aerodynamic design and CFD for the U.S. defense and automotive racing sectors. He earned his Master's Degree in Aerospace Engineering at the University of Florida, following a Bachelor's in the same field from Virginia Polytechnic Institute and State University.

Conor Elrick is an HPC/AI Performance Engineer in the ISG Offerings Group at Lenovo working as part of the HPC/AI Solutions Architect Team. Conor performs setup and performance benchmarking on compute servers, networking infrastructure and storage solutions to push hardware to its limits and get the best results for real world scientific workloads. He has been at Lenovo for over 2 years. Conor earned his PhD in Theoretical Particle Physics from the University of Edinburgh in 2024 and a Master's Degree and Bachelor's Degree in the same field before that, with a focus on physics simulations on compute systems.

Related product families

Product families related to this document are the following:

- ThinkSystem SC777 V4 server

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2470, was created or updated on July 4, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2470
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2470.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
Neptune®
ThinkSystem®

Other company, product, or service names may be trademarks or service marks of others.