

# Dell Validated Design for Technical Computing AI

Dell PowerEdge XE9680 with 8x NVIDIA H100 Tensor Core Accelerators

White Paper

## Abstract

This white paper presents the Dell Validated Design for Technical Computing AI with performance benchmarking results based on the Dell 16G platform with Intel's 5th generation Xeon Scalable Server Processor and NVIDIA (Mellanox) options for the networking fabric.

## Copyright

© 2024 Dell Inc. or its subsidiaries. All rights reserved. Dell Technologies, Dell, and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.

# Contents

|                                  |    |
|----------------------------------|----|
| Introduction.....                | 4  |
| Executive summary.....           | 4  |
| Document purpose.....            | 4  |
| Audience.....                    | 5  |
| Business challenges.....         | 5  |
| Customer challenges.....         | 5  |
| Solution overview.....           | 5  |
| Introduction.....                | 5  |
| Solution architecture.....       | 6  |
| Architecture overview.....       | 6  |
| Management servers.....          | 6  |
| Compute servers.....             | 7  |
| Network design.....              | 7  |
| Storage.....                     | 9  |
| Cluster management software..... | 10 |
| Benchmark disclaimer.....        | 10 |
| Performance.....                 | 10 |
| Introduction.....                | 10 |
| MLPerf.....                      | 10 |
| High-Performance Linpack.....    | 12 |
| HPL-AI.....                      | 14 |
| HPCG.....                        | 14 |
| Conclusion .....                 | 16 |
| Overview.....                    | 16 |
| We value your feedback.....      | 16 |
| References.....                  | 16 |
| Benchmark documentation.....     | 16 |



## Topics:

- [Introduction](#)
- [Business challenges](#)
- [Solution overview](#)
- [Solution architecture](#)
- [Performance](#)
- [Conclusion](#)
- [References](#)

# Introduction

## Executive summary

The Dell Validated Design for Technical Computing (TC) artificial intelligence (AI) serves as a foundation for workload-optimized, rack-level systems. This solution offers flexible design options for compute, networking, and storage components.

The growth of AI applications and their use cases impact nearly all aspects of business and personal life. Generative AI, a branch of AI designed to create data, images, code, and more without explicit human programming, is particularly influential. [Precedence Research](#) reports that the global generative AI market, which was valued at USD 17.65 billion in 2023, is projected to reach USD 803.90 billion by 2033, with a CAGR of 46.5% from 2024 to 2032.

Generative AI applications include:


- Conversational agents and chatbots for customer service
- Natural language interaction and translation
- Audio and visual content creation
- Software programming
- Security, fraud detection, and threat intelligence

Few areas of business and society remain unaffected by this technology. While open-source generative AI models like ChatGPT, Google Gemini, and DALL-E are intriguing, they raise concerns about output ownership, accuracy, truthfulness, and source attribution. As a result, many enterprises are looking to develop their own Large Language Models (LLMs) using proprietary datasets or by fine-tuning existing pretrained models.

## Document purpose

This white paper describes the Dell Validated Design for TC AI, a solution based on the PowerEdge XE9680 server with 8x NVIDIA H100 accelerators and [Dell Omnia](#).

It provides an overview of the key design concepts and a detailed description of the solution architecture, including hardware and software components. It explains how these components are interconnected through the networking design. Additionally, this white paper outlines the validation environment, methodology, and results.

 **NOTE:** The contents of this document are valid for the described software and hardware versions. For information about updated configurations for newer software and hardware versions, contact your Dell Technologies sales representative.

# Audience

This white paper is intended for business leaders such as Chief Technology Officers (CTOs), Chief Information Officers (CIOs), IT infrastructure managers, and systems architects who are involved with or considering the implementation of AI.

## Business challenges

### Customer challenges

The convergence of high-performance computing (HPC), AI, and data analytics has led infrastructure teams to seek HPC systems that support both compute-centric and data-centric uses with a single resource pool. However, some infrastructure teams may prefer HPC clusters that are focused on either traditional HPC workloads for simulation and modeling, or data-centric workloads for AI and data analytics.

Many customers invest significant time and resources in evaluating design choices, including network architecture, storage architecture, file systems, server configurations, CPUs, accelerators, memory, hard drives, operating systems, runtime and user-level libraries, workload managers, applications, and benchmarking. These efforts aim to specify, build, and tune clusters optimized for their needs.

## Solution overview

### Introduction

In the dynamic field of AI and natural language processing, LLMs have become essential for tasks such as text summarization, information retrieval, and content creation. The Meta Llama 3 model, which is known for its ability to comprehend context and produce insightful responses, exemplifies a powerful conversational agent. However, maximizing the performance of these models necessitates strategic, well-designed architecture that includes high-performance servers, high-speed networking, and scalable storage. As organizations strive to harness the full potential of AI, hardware acceleration becomes a pivotal factor in achieving superior performance.

The Dell PowerEdge XE9680 server is the foundation of this collaboration. It offers enterprises unparalleled capabilities with eight NVIDIA H100 Tensor Core GPU accelerators. As the first eight-way GPU platform from Dell, this server enhances application performance by managing the most complex AI, machine learning, deep learning, and HPC workloads.

Table 1. Building block components

| Functional component type   | Optional building block components  |
|-----------------------------|---|
| Infrastructure servers      | AMD: R6625, Intel: R660   |
| Compute servers             | Intel: XE9680   |
| Storage                     | PowerScale F710, Local NVMe 1 TB  |
| Networking                  | <ul style="list-style-type: none"><li>Front-end (access/storage) network: Dell PowerSwitch Z9432F-ON</li><li>Back-end GPU network: NVIDIA QM9790 NDR</li><li>OOB network: Dell PowerSwitch N3248TE-ON</li></ul> |
| Cluster management software | Omnia 1.6   |

# Solution architecture

## Architecture overview

The Dell Validated Design for TC AI addresses the challenges of customizing LLMs for enterprise use cases. While LLMs demonstrate tremendous potential in natural language processing tasks, they require specialized infrastructure for efficient customization and deployment.

This reference architecture provides organizations with guidelines and best practices to design and implement scalable, efficient, and reliable infrastructure tailored for training and customizing generative AI models. While its primary focus is on LLM customization, the architecture can also be adapted for training discriminative or predictive AI models.

The following figure illustrates the key components of the reference architecture:

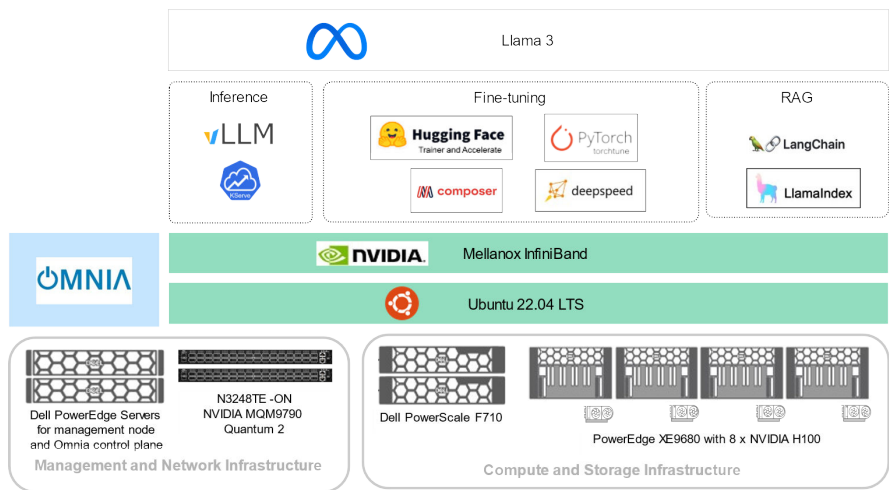


Figure 1. High-level solution architecture

## Management servers

Infrastructure servers provide non-compute services in the cluster, including administration and user access. These servers typically fall into two roles: management and login nodes. The exact configuration and number of infrastructure servers depend on the cluster size and requirements. Although login servers are not required, separating users from critical management systems simplifies administration and minimizes unplanned downtime. For example, a typical system has one login server for every 30 to 100 users. Infrastructure nodes can also provide additional services, such as NFS storage service.

The baseline configurations for head and control plane nodes are based on the PowerEdge R760. Density is not a concern because management nodes constitute a small fraction of the overall cluster, allowing the use of regular 1U or 2U systems. Typically, clusters have matching platform architectures for both infrastructure and compute servers to ease administration. Login nodes specifically benefit from being of the same architecture as application nodes.

The following table provides the recommended minimum configuration for the management head node and the control plane node:

Table 2. PowerEdge R760 head node and control plane configuration

| Component        | Head node and control plane nodes     |
|------------------|---------------------------------------|
| Server model     | 2x PowerEdge R760                     |
| CPU              | 2x Intel Xeon Platinum 8468 Processor |
| Memory           | 512 GB, 16x 32 GB 4800 MT/s           |
| Operating system | Ubuntu 22.04                          |
| RAID controller  | PERC H755 RAID 6                      |
| Storage          | Local: 10x 960 GB SATA                |

**Table 2. PowerEdge R760 head node and control plane configuration (continued)**

| Component | Head node and control plane nodes    |
|-----------|--------------------------------------|
| Network   | Broadcom BCM57414 NetExtreme-E 10 GB |

Consider the following recommendations for configuring the Omnia control plane:

- A dedicated PowerEdge server for Omnia is recommended, as the deployment of Omnia requires a control plane.
- Since the Omnia node does not require heavy computing, a single-processor server is sufficient, although a dual-processor server was used for this study.
- Omnia generally uses NFS, minimizing the need for local storage. However, you can customize the server configuration for additional local user storage space by expanding the drive quantity and RAID type. NVMe drives are available for higher performance.
- This release of Omnia does not support installing and configuring Slurm as part of the deployment process. For this project, PMix and Slurm were compiled and installed from source.

## Compute servers

The compute infrastructure is a critical component of the design, ensuring the efficient implementation of AI models. The PowerEdge XE9680 server is a two-socket, 6U server that supports eight NVIDIA H100 accelerators, offering more options for AI performance.

In this design, PowerEdge XE9680 servers are configured as worker nodes in a cluster. Omnia, an open - source software for deploying and managing clusters, is used to deploy operating systems, while other required software stacks are configured manually.

The following table provides a recommended configuration for a PowerEdge XE9680 GPU compute node:

**Table 3. PowerEdge XE9680 GPU compute node**

| Component         | Compute Nodes  |
|-------------------|--|
| Server model      | 4x PowerEdge XE9680  |
| CPU               | 2x Intel Xeon Platinum 8480+ Processor (105M cache, 2.00 GHz)  |
| Memory            | 2 TB, 32x 64 GB 4800 MTs   |
| Operating system  | Ubuntu 22.04   |
| Storage           | <ul style="list-style-type: none"> <li>• Local: 1.92 TB NVMe MZ-WLR3T8B</li> <li>• Shared: 103 TB NFS Mount PowerScale F710</li> </ul>                     |
| Networking        | <ul style="list-style-type: none"> <li>• GPU network: 8x NVIDIA Mellanox NDR 400</li> <li>• Storage network: 1x NVIDIA CX6 port set to ETH mode</li> </ul> |
| GPU (accelerator) | 8x NVIDIA H100 SXM 80 GB Tensor Core GPU   |

The CPU memory allocation in the PowerEdge XE9680 GPU compute node configuration must exceed the combined GPU memory footprint. Therefore, we recommend a minimum of 2 TB of total RAM. While LLM tasks primarily rely on GPUs and do not significantly tax the CPU and memory, it is advisable to equip the system with high-performance CPUs and larger memory capacities. This provisioning ensures sufficient capacity for various data processing activities, machine learning operations, and monitoring and logging tasks. The objective is to ensure that the servers provide ample CPU and memory resources for these functions, preventing any potential disruptions to the critical AI operations on the GPUs.

## Network design

Dell Technologies Open Networking is designed around a highly scalable, cloud-ready data center network fabric. It uses the Dell Enterprise SONiC network operating system. IT organizations can leverage SONiC's first commercial offering for innovation, automation, and reliability, with enterprise enhancements and global support for cloud, data center, and edge fabrics.

Dell Technologies has long been a pioneer in open networking, and its leadership in the development and deployment of SONiC demonstrates its continued commitment to community innovation, collaboration, and contribution. Dell's approach to SONiC is

built on the principles of openness and interoperability. This empowers enterprises to avoid vendor lock-in, enabling them to customize and optimize their networks to meet specific needs.

Enterprise SONiC Distribution by Dell Technologies offers several features that are designed to enhance the performance, efficiency, and connectivity of AI fabrics. Enterprise SONiC brings substantial advancements in AI fabric enablement. With features such as Dynamic Load Balancing with Adaptive Routing and Enhanced User-Defined Hashing, this release empowers organizations to use AI fabrics more effectively. Dynamic Load Balancing ensures optimal use of links in an AI fabric, while Adaptive Routing enhances forwarding behavior, maximizing the performance and efficiency of network resources. These advanced network architecture capabilities allow AI data flow to simultaneously access all available paths to its destination.

Deploying AI technologies presents challenges, such as technical complexities, shortages of skilled professionals, and the limitations of proprietary technologies like InfiniBand. These limitations complicate integration, leading to high costs, long evaluation periods, and vendor lock-in.

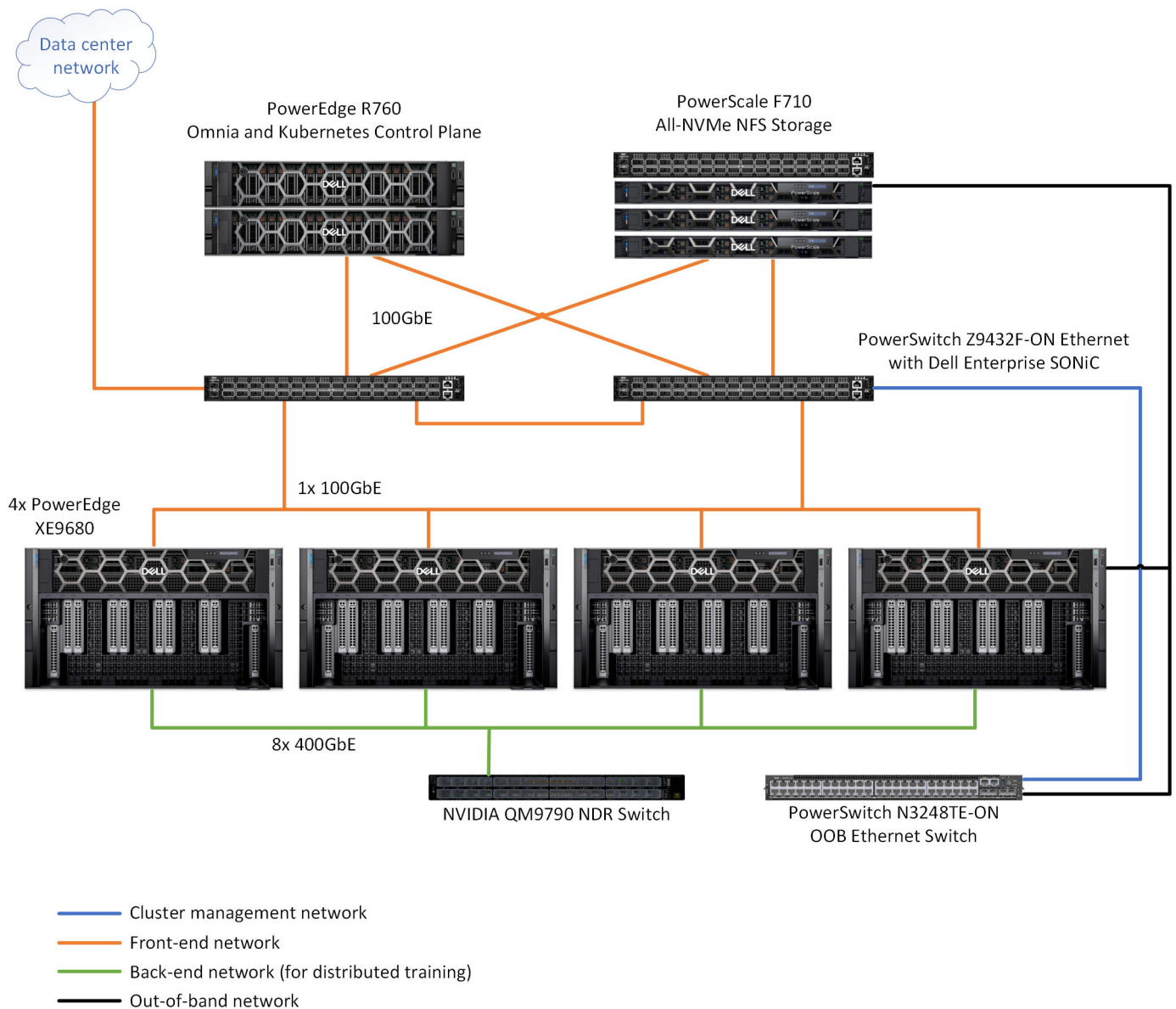
As AI models scale from billions to trillions of parameters, the need for substantial data transfer becomes critical, and any network-induced delay can affect performance. Therefore, rapid bulk data transfer with elephant flows from source to destination is vital for job completion. Although per packet latency is important, the total time that is required to complete an entire processing step is even more crucial. These large AI clusters also require coordinated congestion management to avoid packet loss and ensure efficient GPU utilization, along with synchronized management and monitoring to optimize both compute and network resources. This design incorporates three physical networks:

1. Frontend network for management, storage, client/server traffic powered by two PowerSwitch Z9432F-ON
2. Backend network for internode GPU communication powered by NVIDIA Quantum-2 QM9790
3. Out-of-band traffic management powered by PowerSwitch N3248TE-ON

The PowerSwitch Z-series switches are high-performance, open, and scalable data center switches used for spine, core, and aggregation applications. The PowerSwitch Z9432F-ON is a high-density 400 GbE fabric switch, offering up to 32 ports of 400 GbE or up to 128 ports of 100 GbE with breakout cables. It provides a broad range of functionalities to meet the growing demands of today's data center environments. The PowerSwitch Z9664F-ON offers high density with either 64 ports of 400 GbE in a QSFP56-DD form factor or 256 ports of 100 GbE in a 2U design. It can function as a 10, 25, 40, 50, 100, or 200 switch with breakout cables, supporting up to 256 ports.

The following figure shows the network architecture, which shows the network connectivity for compute servers:





**Figure 2. Network design**

## Storage


Dell PowerScale offers massive AI performance with ultimate density, accelerating all phases of the AI pipeline, from model training to inferencing and fine-tuning. With up to 24 NVMe SSD drives per node and 300 PBs of storage per cluster, it ensures optimal GPU utilization for large-scale model training and drives faster time to AI insights with up to 127% improved throughput.

**NOTE:** Based on internal testing, comparing streaming write of F910 on OneFS 9.8 to streaming write of F900 on OneFS 9.5. Results might vary. (April 2024)

PowerScale is designed for AI-optimized infrastructure. It was one of the first storage products to offer low latency storage access with Network File System over Remote Direct Memory Access (NFS over RDMA), multitenant capabilities, simultaneous multiprotocol support, and 6x9s availability and resiliency to ensure uninterrupted uptime.

Our next-generation all-flash systems build on that foundation, leveraging continuous software and hardware innovation to form a key component of Dell's AI-ready data platform. They offer:

- Multicloud agility with Dell APEX File Storage for public cloud portfolio
- Federal-grade security features to safeguard the AI process from attacks such as data poisoning and model inversion
- Exceptional efficiency with scale-out NAS to manage AI data growth while controlling storage costs

 **NOTE:** Based on Dell analysis comparing efficiency-related features including data reduction, storage capacity, data protection, hardware, space, life cycle management efficiency, and ENERGY STAR certified configurations. (June 2023)

Continuous PowerScale innovation extends into the AI era with the introduction of the next generation of PowerEdge-based nodes, including the PowerScale F710. The new PowerScale all-flash nodes use Dell PowerEdge 16G servers, unlocking the next generation of performance. Regarding software, the F710 takes advantage of significant performance improvements in PowerScale OneFS 9.7. Combining the latest hardware and software innovations, the F710 can tackle the most demanding workloads with ease.

## Cluster management software

[Dell Omnia](#) is an open-source software that is made to deploy and manage high-performance clusters for HPC, AI, and data analytics workloads. Omnia installs Kubernetes and Slurm to manage jobs and supports the installation of many other packages and services for diverse workloads on a converged solution. Developers continually extend Omnia to expedite the deployment of new infrastructure into resource pools, which can be allocated and reallocated to different workloads. Omnia simplifies and accelerates IT's ability to provide the right tools for the right job on the right infrastructure at the right time.

## Benchmark disclaimer

Benchmark results depend on workload, specific application requirements, and system design and implementation. Relative system performance varies based on these and other factors. Benchmarking results should not replace specific customer application benchmarks for critical capacity planning or product evaluation decisions.

All performance data in this report was obtained in a rigorously controlled environment. Results that are obtained in other operating environments may vary significantly. Dell Technologies does not warrant or represent that users can or will achieve similar performance results.

# Performance

## Introduction

Benchmarking an infrastructure before using it for LLM tasks, such as training and inference, is crucial for several reasons. First, it helps understand the infrastructure's capacity to handle the computational demands of these tasks. Many AI workloads, such as generative AI or LLM tasks, require significant computational resources, and benchmarking can provide insights into whether the current infrastructure can meet these demands. Second, it allows for the optimization of resource allocation, ensuring efficient and cost-effective use of the infrastructure. Third, benchmarking can identify potential bottlenecks that may hinder the AI task performance. By addressing these issues early, you can ensure smooth and efficient operation. Lastly, benchmarking provides a baseline for measuring future upgrades or changes to the infrastructure, aiding continuous improvement efforts. Therefore, benchmarking is a vital step in preparing an infrastructure.

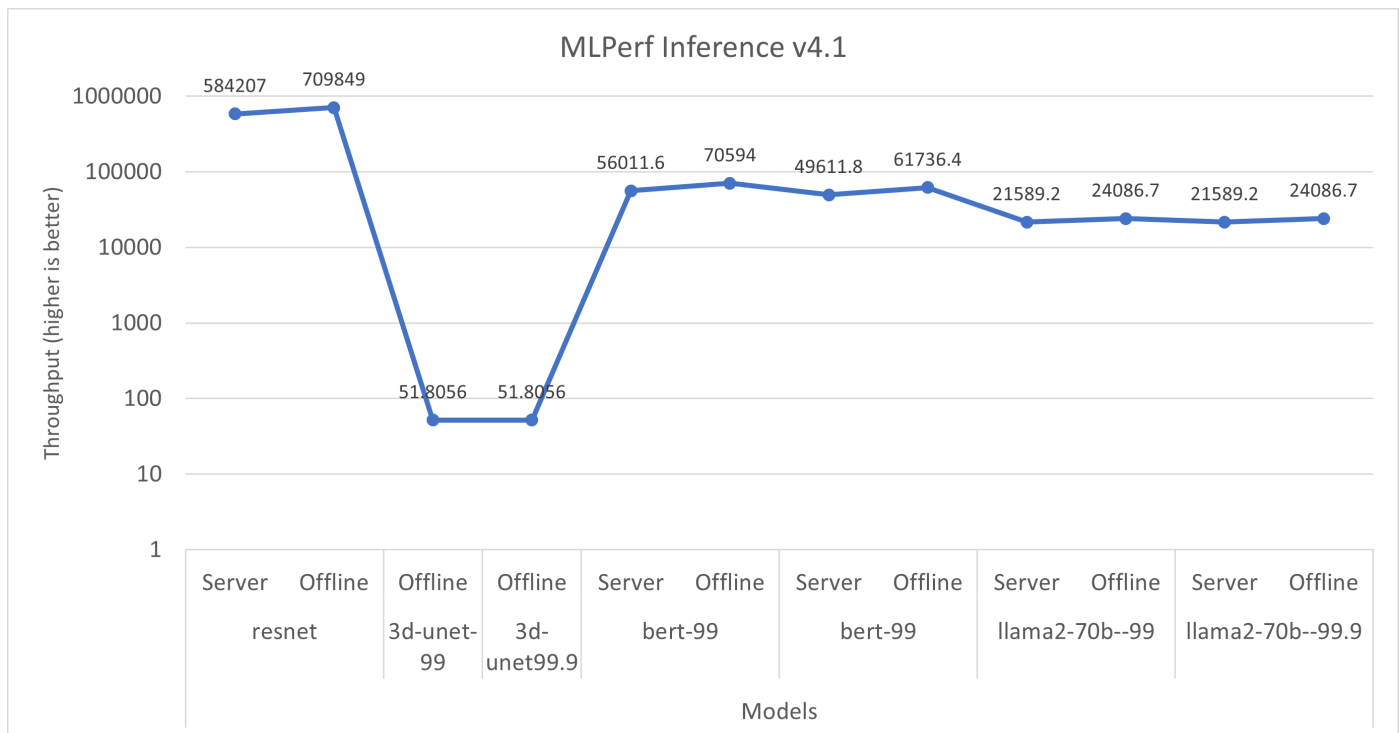
Four specific applications were tested on the proposed solution: [MLPerf Training v4.0](#) and [Inference v4.1](#) and [NVIDIA HPC-Benchmarks 24.06](#) for HPL, HPL-AI, and HPCG.

## MLPerf

MLPerf is a benchmark suite that is used to evaluate training and inference performance of on-premises and cloud platforms. It serves as an independent, objective performance yardstick for software frameworks, hardware platforms, and cloud platforms in machine learning. Developed and continuously evolved by a consortium of AI community researchers and developers, MLPerf aims to provide developers with a tool to evaluate hardware architectures and the diverse range of advancing machine learning frameworks.

MLPerf Inference benchmarks measure the speed at which a trained neural network can perform inference tasks.

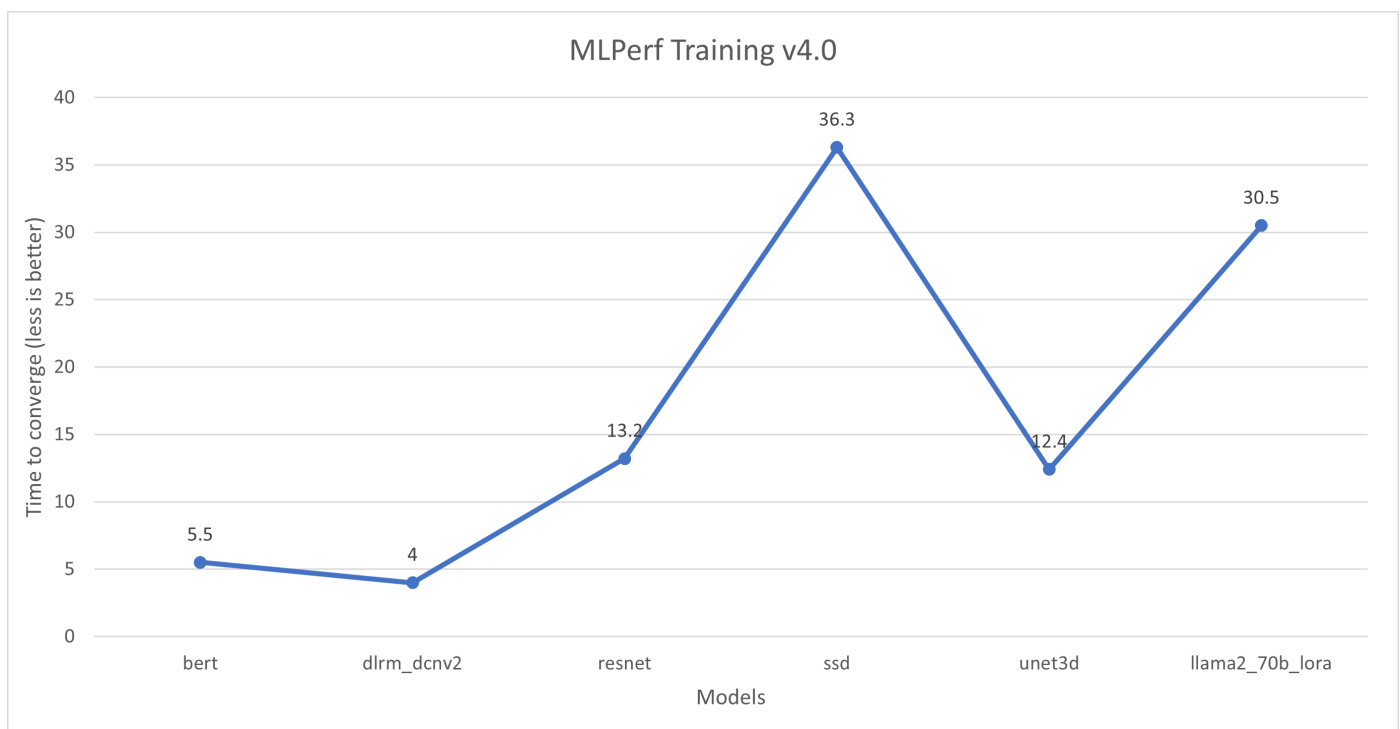
Code can be found on [GitHub](#).



**Figure 3. MLPerf Inference v4.1 results for a single XE9680 with 8x H100-SXM-80GB GPUs**

The MLPerf Training benchmarking suite measures the time that is required to train machine learning models to a target level of accuracy on new data.

Code can be found on [GitHub](#).



**Figure 4. MLPerf Training v4.0 results for a single XE9680 with 8x H100-SXM-80GB GPUs**

# High-Performance Linpack

High Performance Linpack ([HPL](#)) is a classic HPC benchmark suite that is used to measure system size and performance. The HPL benchmark performs a LU factorization on a large matrix to solve a uniformly random system of linear equations and reports a floating-point execution rate using a standard formula for operation count . The NVIDIA HPL benchmark uses double precision (64-bit) arithmetic. NVIDIA's HPL benchmark is intended for distributed-memory computers that are equipped with NVIDIA GPUs and is based on the Netlib HPL. For more information about Dell Technology's latest large-scale benchmarking publications, go to [top500.org](#).

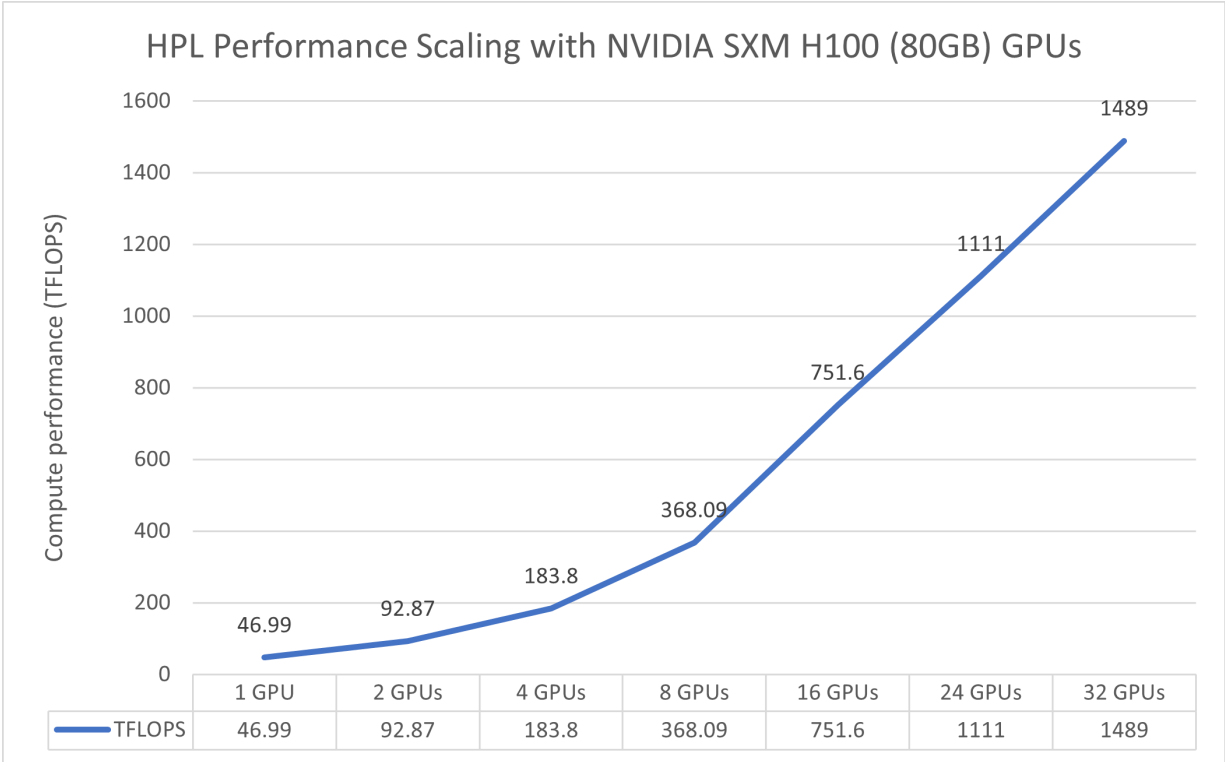


Figure 5. Performance scaling of NVIDIA's HPL benchmark with 1, 2, 4, 8, 16, 24, 32 H100 SXM 80GB GPUs

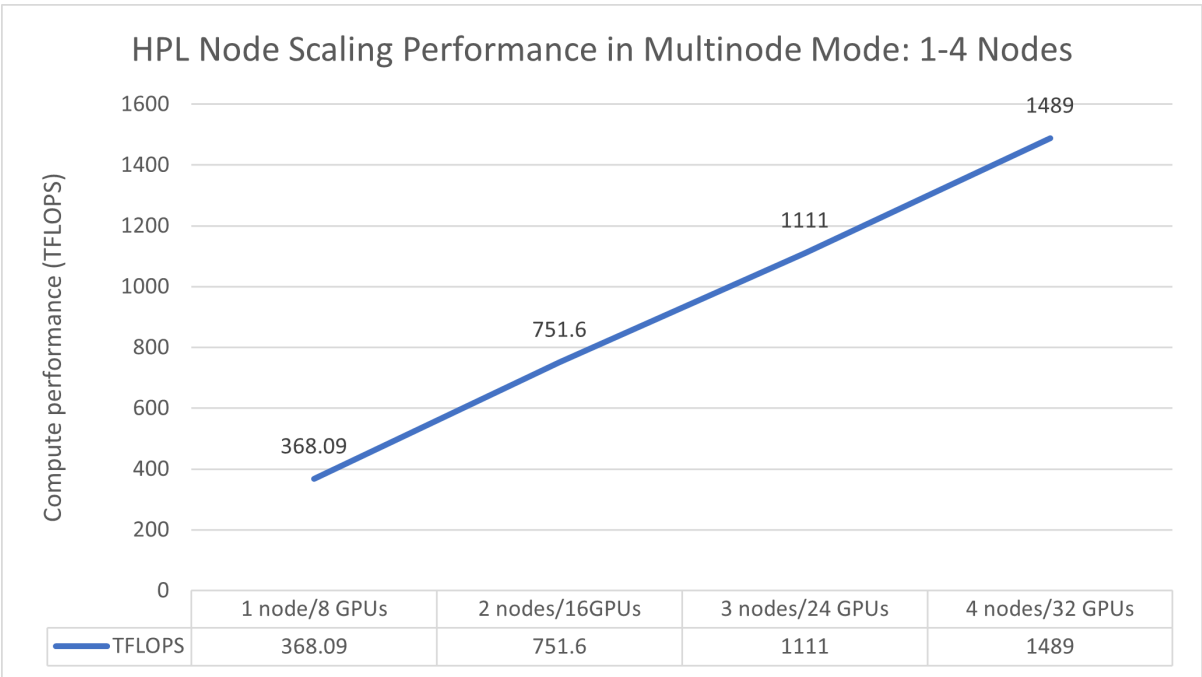
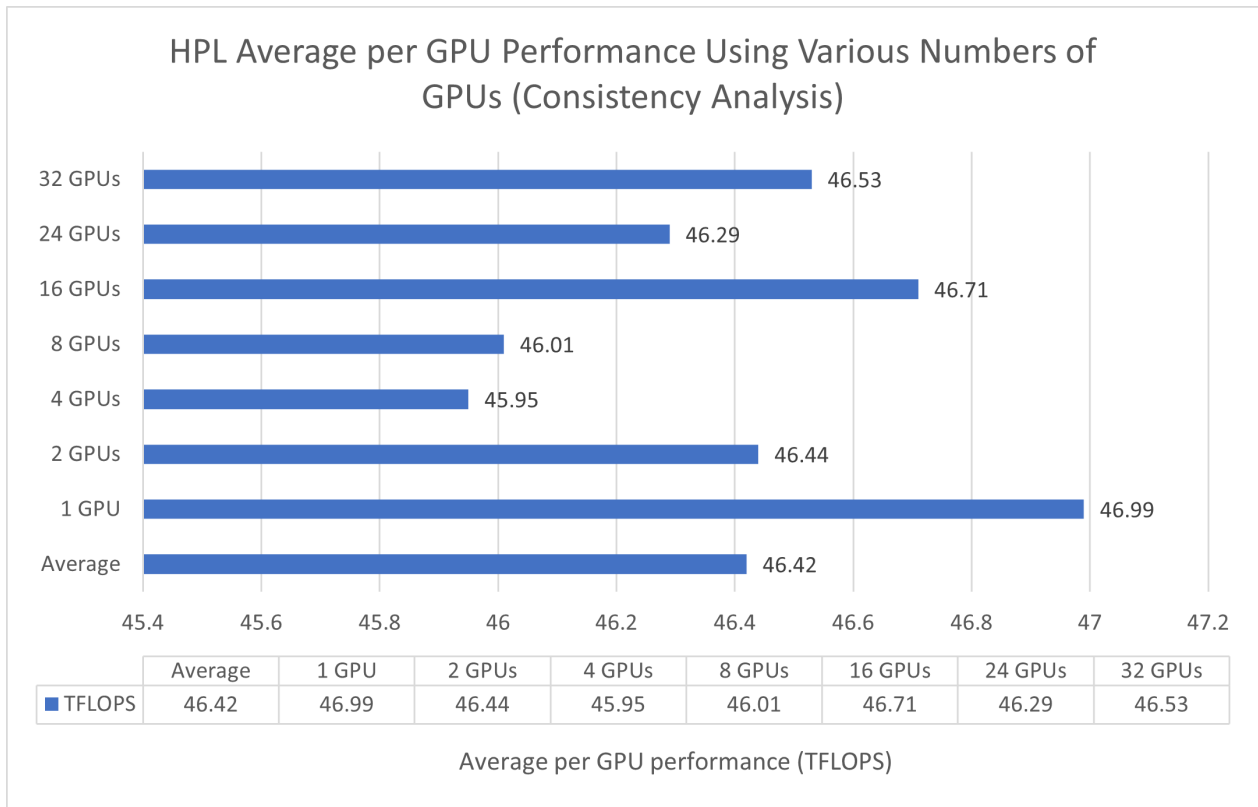
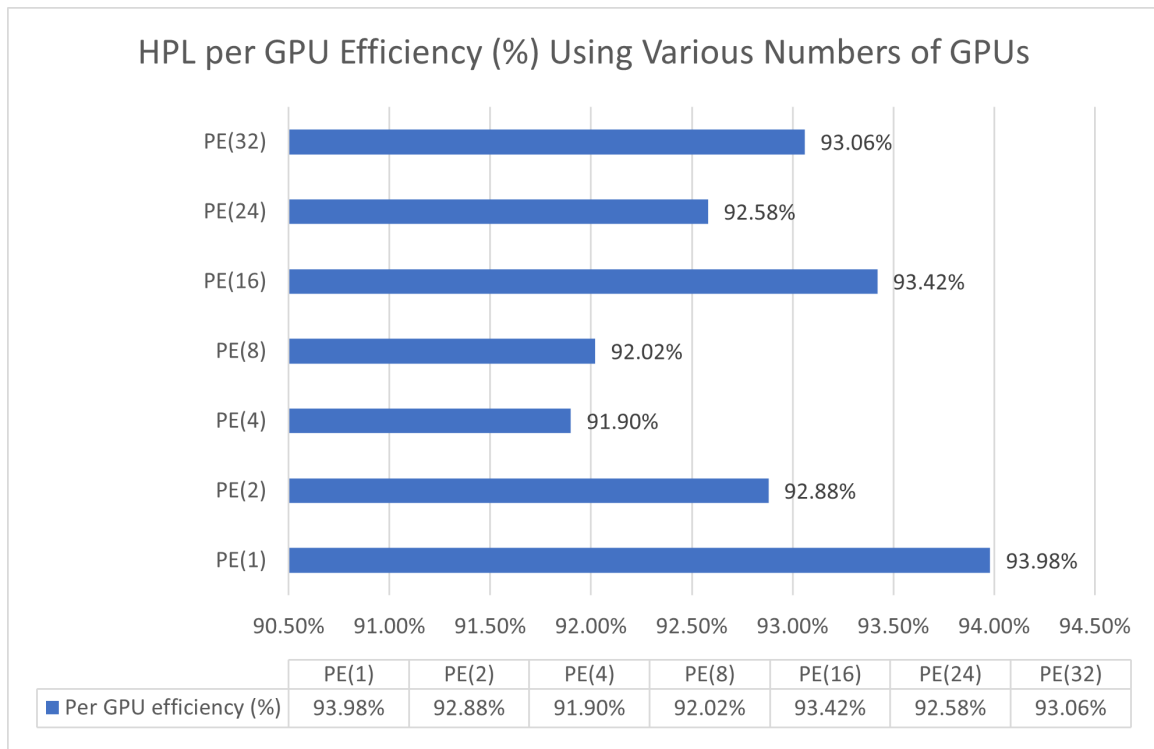


Figure 6. HPL scaling performance using 1, 2, 3, 4 Dell XE9680 servers in multinode mode



**Figure 7. NVIDIA HPL average per GPU compute performance analysis using 1, 2, 4, 8, 16, 24, 32 GPUs**



**Figure 8. Performance measurement of per GPU efficiency (PE) using various numbers of NVIDIA's H100 SXM 80 GB GPUs ( $PE(n) = ((\text{per GPU computer performance}) / (\text{target GPU performance})) * 100\%$ )**

# HPL-AI

HPL-MxP (also known as HPL-AI) benchmark demonstrates the convergence of HPC and AI workloads by solving a system of linear equations using innovative mixed-precision algorithms. See [HPL-MxP](#) for more information.

HPL-MxP performance improves with larger matrix sizes, optimizing the use of computational resources. In most cases, the benchmark is configured to maximize GPU memory utilization by adjusting the global matrix size. Fine-tuning these parameters depends on the specific hardware setup. For instance, on NVIDIA GPUs, adjusting matrix dimensions, block size, and process grid dimensions is crucial to achieving optimal performance tailored to the system's architecture.

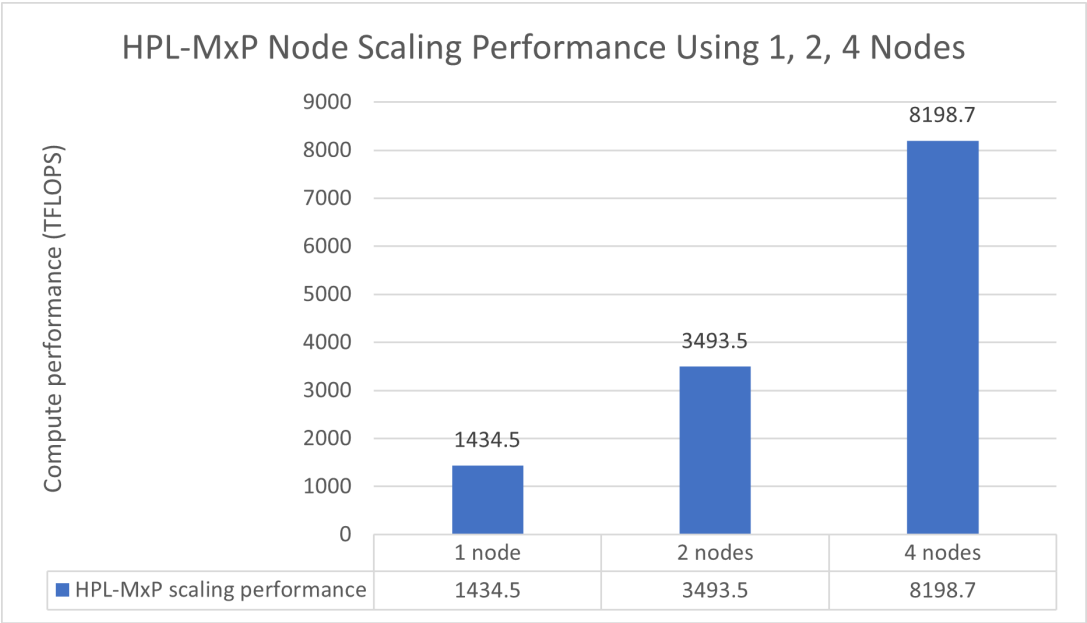


Figure 9. HPL-MxP Node Scaling Performance Using 1, 2, 4 Nodes

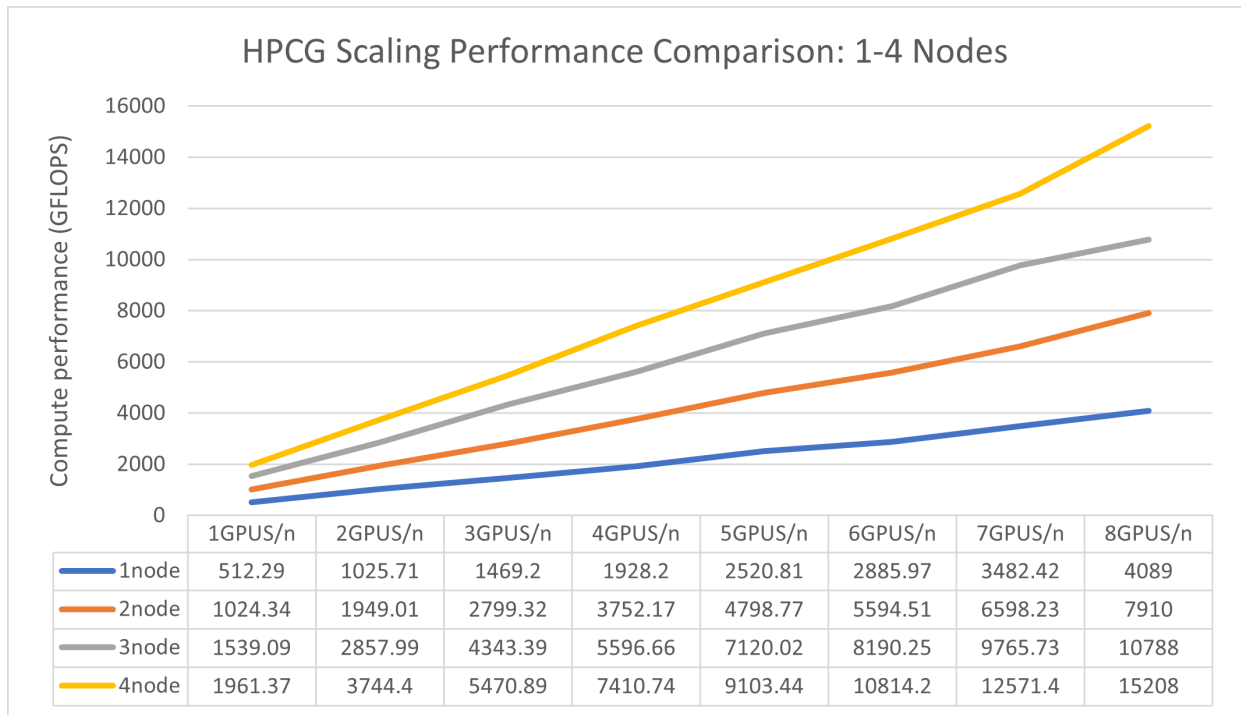
Table 4. Speedup comparison between mixed-precision workload and HPC FP64 precision workload

| Benchmark (TFLOPS)   | One node | Two nodes | Four nodes |
|----------------------|----------|-----------|------------|
| NVIDIA HPL           | 368      | 752       | 1489       |
| NVIDIA HPL-MxP       | 1435     | 3494      | 8199       |
| SPEEDUP: HPL-MxP/HPL | 3.90     | 4.65      | 5.51       |

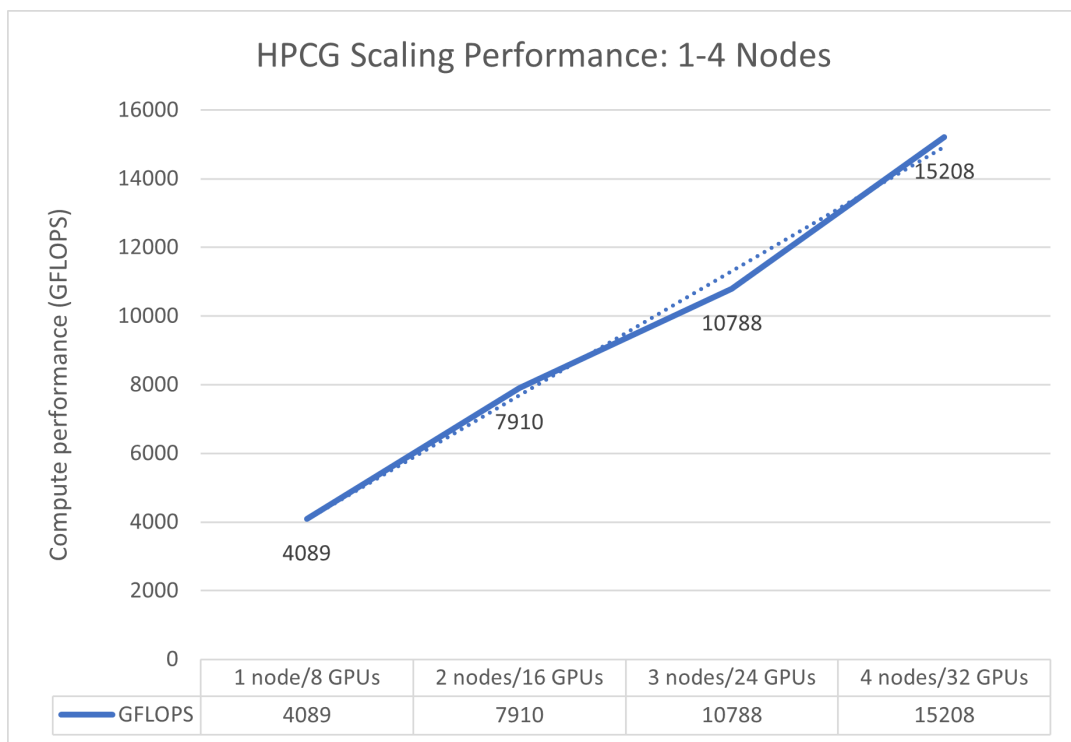
The HPL-MxP benchmark integrates HPC and AI workloads by solving systems of linear equations using innovative mixed-precision algorithms. This benchmark provides a more realistic representation of mixed-precision tasks for modern deep learning and large language models. Table 4 illustrates the speedup that is achieved when transitioning from a pure 64-bit (FP64) workload to the HPL-MxP mixed-precision workload.

# HPCG

High Performance Conjugate Gradients (HPCG) complements HPL by incorporating data access and compute patterns that closely match current HPC workloads.



**Figure 10. Breakdown of performance measurements for GPU scaling and node scaling cases, demonstrating near-linear scaling results**



**Figure 11. Near-linear scaling of compute performance measured by NVIDIA's HPCG benchmark using 1, 2, 3, and 4 Dell XE9860 servers (each equipped with 8 NVIDIA H100 SXM 80GB GPUs)**

# Conclusion

## Overview

The Dell Validated Design for TC AI with NVIDIA H100 Accelerators addresses the needs of enterprises developing and running custom AI models using domain-specific data relevant to their own organization.

Dell Technologies has designed a scalable, modular, and high-performance architecture that enables enterprises to quickly design and deploy AI inferencing solutions that are customized to their specific needs using fine-tuning and RAG methodologies. These solutions have been validated and performance-tested to accelerate time to value and reduce risk and uncertainty through a proven design.

Dell Technologies enables organizations to deliver full-stack AI solutions that are built on the best of Dell infrastructure and software, which is combined with NVIDIA accelerators, AI software, and AI expertise. This combination enables enterprises to use purpose-built generative AI on-premises to solve their business challenges. Together, we are leading the next wave of innovation in the enterprise AI landscape.

## We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and solution documentation. Contact the Dell Technologies Solutions team by [email](#).

# References

## Benchmark documentation

The following list includes links to the benchmarks used in this white paper.

- [MLPerf](#)
- [HPL](#)
- [HPL-AI](#)
- [HPCG](#)