

Hitachi iQ with NVIDIA HGX

Reference Architecture Guide

© 2024 Hitachi Vantara LLC. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including copying and recording, or stored in a database or retrieval system for commercial purposes without the express written permission of Hitachi, Ltd., Hitachi Vantara, Ltd., or Hitachi Vantara Corporation (collectively "Hitachi"). Licensee may make copies of the Materials provided that any such copy is: (i) created as an essential step in utilization of the Software as licensed and is used in no other manner; or (ii) used for archival purposes. Licensee may not make any other copies of the Materials. "Materials" mean text, data, photographs, graphics, audio, video and documents.

Hitachi reserves the right to make changes to this Material at any time without notice and assumes no responsibility for its use. The Materials contain the most current information available at the time of publication.

Some of the features described in the Materials might not be currently available. Refer to the most recent product announcement for information about feature and product availability, or contact Hitachi Vantara LLC at https://support.hitachivantara.com/en_us/contact-us.html.

Notice: Hitachi products and services can be ordered only under the terms and conditions of the applicable Hitachi agreements. The use of Hitachi products is governed by the terms of your agreements with Hitachi Vantara LLC.

By using this software, you agree that you are responsible for:

1. Acquiring the relevant consents as may be required under local privacy laws or otherwise from authorized employees and other individuals; and
2. Verifying that your data continues to be held, retrieved, deleted, or otherwise processed in accordance with relevant laws.

Notice on Export Controls. The technical data and technology inherent in this Document may be subject to U.S. export control laws, including the U.S. Export Administration Act and its associated regulations, and may be subject to export or import regulations in other countries. Reader agrees to comply strictly with all such regulations and acknowledges that Reader has the responsibility to obtain licenses to export, re-export, or import the Document and any Compliant Products.

Hitachi and Lumada are trademarks or registered trademarks of Hitachi, Ltd., in the United States and other countries.

AIX, DB2, DS6000, DS8000, Enterprise Storage Server, eServer, FICON, FlashCopy, GDPS, HyperSwap, IBM, OS/390, PowerHA, PowerPC, S/390, System z9, System z10, Tivoli, z/OS, z9, z10, z13, z14, z15, z16, z/VM, and z/VSE are registered trademarks or trademarks of International Business Machines Corporation.

Active Directory, ActiveX, Bing, Excel, Hyper-V, Internet Explorer, the Internet Explorer logo, Microsoft, Microsoft Edge, the Microsoft corporate logo, the Microsoft Edge logo, MS-DOS, Outlook, PowerPoint, SharePoint, Silverlight, SmartScreen, SQL Server, Visual Basic, Visual C++, Visual Studio, Windows, the Windows logo, Windows Azure, Windows PowerShell, Windows Server, the Windows start button, and Windows Vista are registered trademarks or trademarks of Microsoft Corporation. Microsoft product screen shots are reprinted with permission from Microsoft Corporation.

All other trademarks, service marks, and company names in this document or website are properties of their respective owners.

Copyright and license information for third-party and open source software used in Hitachi Vantara products can be found in the product documentation, at <https://www.hitachivantara.com/en-us/company/legal.html>.

Feedback

Please send comments to doc.feedback@hitachivantara.com. Include the document title and number, including the revision level (for example, -07), and refer to specific sections and paragraphs whenever possible. All comments become the property of Hitachi Vantara LLC.

Thank you!

Revision history

Changes	Date
Initial release	November 2024

Reference Architecture Guide

Purpose

AI and Generative AI are revolutionizing industries by enhancing capabilities in art creation, fraud detection, and even software development. Organizations across various sectors recognize the transformative potential of AI, appreciating its significant value in shaping market dynamics and driving business success. From graphic design and code generation to crafting marketing slogans, AI applications are not only versatile but are proving to be integral in delivering immediate benefits to businesses. As these technologies continue to evolve, the possibilities for innovation and efficiency are limitless.

Organizations are constantly looking for ways to automate their business, accelerate time-to-market and develop new insights, products, or innovations to propel their business forward. Hitachi iQ allows organizations to realize these goals for their business through intelligent, performant, scalable, and flexible AI and GenAI solutions.

Whether you are looking for industry-specific AI solutions or just starting to identify general-purpose capabilities, Hitachi iQ has the power to automate your business processes and improve your customer experience.

Industry specific outcomes

Unlike other approaches on the market, Hitachi iQ goes beyond the basics of storage and infrastructure by layering industry-specific AI outcomes for industries such as finance, energy, transportation, manufacturing, and more. These solutions provide relevance and simplification to customers by accelerating the development and adoption of AI solutions and services into their ecosystem.

The following are additional use cases powered by NVIDIA and Hitachi iQ:

- Health and Life Sciences (medical devices, genomics, drug discovery, smart hospitals)
- Automotive (parts design, autonomous vehicles)
- Telecommunications (network security, 5G network planning)
- Large Language Models (training, fine-tuning)

From entry level to enterprise capacity, Hitachi iQ is a versatile solution for AI needs.

This reference architecture focuses on the Hitachi iQ enterprise solution with NVIDIA H100 Tensor Core GPUs powered by Hitachi Content Software for File.

The intended audience of this document is:

- Data scientists and data engineers
- AI developers and architects
- Data analysts
- Security operations

- Storage administrators
- System administrators
- IT professionals

This technical paper assumes that you are already familiar with the following:

- GPU-based server products
- General storage concepts
- Common IT storage practices
- Kubernetes Orchestration Platform
- InfiniBand and general networking concepts

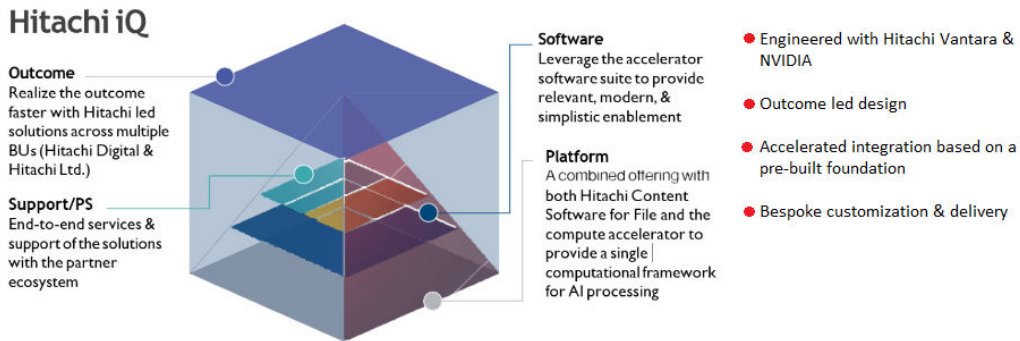
Solution overview

Unlike conventional AI offerings, the Hitachi Vantara AI portfolio, Hitachi iQ, transcends basic integration and storage capabilities by layering industry-specific AI outcomes within the AI solution. This approach ensures that outcomes are finely tuned to the unique needs and objectives of each organization.

Hitachi iQ is an accelerated solution that provides unified access to data, irrespective of where it resides, while ensuring explainability, lineage, data accuracy, security, and traceability at any given point for mission-critical solutions. By creating these optimized AI deployments with end-to-end software-defined AI and data analytics capabilities, Hitachi iQ can streamline the development and deployment of production-grade AI applications from pilot to production.

At the heart of Hitachi iQ is the data platform, Hitachi Content Software for File, a highly scalable, distributed file system designed to work seamlessly with Hitachi Vantara's object storage portfolio. As a software-defined solution, Content Software for File is a software-defined solution that has been designed to be deployed on-premises, in the cloud, or in a hybrid cloud environment, supporting a variety of high-performance applications and workflows including AI, ML, and traditional analytics. With the capability to support over 10 Exabytes, a single cluster can meet the extensive data requirements of any modern organization's data platform needs.

With the Hitachi iQ portfolio, customers will benefit from flexible infrastructure that has been specifically engineered to provide the highest performing AI platform, while simplifying the enterprise AI journey, leading to improved time to value, better economics at scale, and the flexibility required to scale to meet the demands of the enterprise AI workload, regardless of the size, all while being backed by the Hitachi Vantara award-winning support and services organizations.



This reference architecture demonstrates a BasePOD reference configuration of Hitachi iQ Enterprise Solution with NVIDIA HGX™ H100 and Hitachi Content Software for File (HCSF) storage.



Key components and technologies

Hitachi Content Software for File

The unique architecture of Hitachi Content Software for File is different from legacy storage systems, storage appliances, and hypervisor-based software-defined storage solutions as it not only overcomes traditional storage scaling and file sharing limitations but also allows unified file access via POSIX, NFS, SMB, S3 and NVIDIA® GPUDirect® Storage. Hitachi Content Software for File provides a rich enterprise feature set, including local snapshot and remote snapshot offload, automated tiering, dynamic rebalancing, backup, filesystem encryption, authentication integration, quotas, and much more.

The following are Hitachi Content Software for File Benefits:

- Highest performance across all IO profiles — ideal for mixed workloads, including heavy metadata operations
- Scalable capacity — start as small as 207TB and scale to hundreds of petabytes in a single namespace
- Strong security — keep data safe from threat or rogue actors with both software and hardware encryption
- Hybrid Cloud — burst to all the major cloud providers for compute agility or run natively in the cloud
- Backup — offload backups straight to HCP, HCP Cloud Scale or public cloud for long term retention
- Best economics — combine flash and disk for best cost at scale

What sets Hitachi iQ apart is Hitachi Content Software for File, a fully distributed parallel file system that delivers the highest performance file services by leveraging NVMe flash. Also included is integrated tiering that seamlessly expands the namespace to and from hard disk drive (HDD) object storage, without the need for special data migration software or complex scripts; all data resides in a single namespace for easy access and management.

Hardware capabilities

A single file system can support billions of directories and trillions of files, delivering a scalability model more akin to object stores than NAS systems, and directories scale with no loss in performance. Hitachi Content Software for File supports up to 1024 files systems, and up to 24,000 snapshots in a single cluster.

- 6.4 trillion files or directories
- 14 exabytes of managed capacity in the global namespace
- 6.4 billion files in a directory
- 4 petabytes for a single file

The following networking technologies are supported:

- NVIDIA Quantum InfiniBand — 400 Gb/s (NDR400), 200 Gb/s (NDR200, HDR), and 100 Gb/s (EDR)
- NVIDIA Spectrum Ethernet — 100 Gb/s, 200 Gb/s, and 400 Gb/s



Note: This architecture uses specific vendors that are required to achieve performance SLAs. However, multiple vendors offer a range of products that support Ethernet at varying speeds, including 100 Gb/s, 200 Gb/s, and 400 Gb/s, catering to different network infrastructure needs and requirements. Contact a Hitachi Vantara technical sales specialist to discuss which products suit your needs.

Software capabilities

Point-and-click simplicity allows users to rapidly provision new storage; create and expand file systems within a global namespace, establish tiering policy, data protection, encryption, authentication, permissions, NFS, SMB and S3 configuration, read-only or read-write snapshots, snapshot-to-objects, and quality of service policies, as well as monitor overall system health. Detailed event logging provides users the ability to view system events and status over time or drill down into event details with point-in-time precision via the time-series graphing function.

Hitachi Content Software for File has a built-in, policy-based automated data management feature, that transparently moves data across storage types according to the data temperature. Hitachi Content Software for File supports moving data from the NVMe flash storage tier to HCP and HCP Cloud Scale or cloud-based object storage. Snapshots can be saved to lower-cost cold storage such HCP and HCP Cloud Scale object storage.

Hitachi Content Software for File supports user-definable snapshots for routine data protection including backup. For example, snapshots can be used to back up files locally on the flash tier as well as making copies to cloud storage tiers for backup or disaster recovery.

The following protocols are supported by the solution:

- POSIX Compliant Client
- NVIDIA® Magnum IO™ GPUDirect® Storage (GDS)
- NFS (Network File System) v3 and v4.1
- SMB (Server Message Block) v2 and v3
- S3 (Simple Storage Service)

Hitachi Content Platform

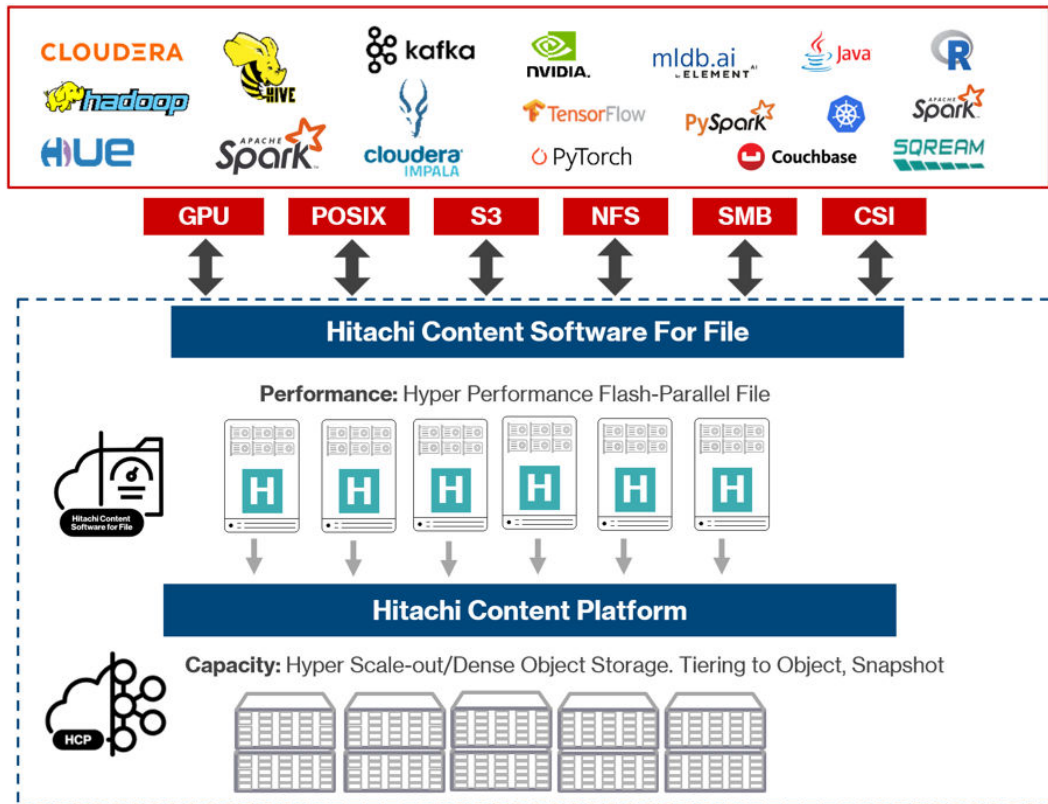
Hitachi Content Platform (HCP) is a secure, simple, and intelligent web-scale object storage platform that delivers superior scale, performance, security, efficiency, and interoperability. It allows any organization to deliver unique, feature-rich, private, hybrid, multi-cloud, or public cloud storage services at a cost comparable to public cloud. The rich feature set and extensive ecosystem surrounding the platform allow organizations to improve efficiencies and optimize costs. They can choose to move data to on-premises storage tiers, off-site to a choice of public cloud providers or to a combination of both.

Hitachi iQ provides additional benefits by connecting Hitachi Content Software for File with Hitachi Content Platform (HCP) object storage for data management, protection, and governance capabilities. Hitachi Content Software for File and HCP are a tightly coupled solution stack that delivers an appliance-like experience with a single, unified namespace across NVMe and object storage for 'always hot' access to data. The distributed storage design ensures that there are no hotspots in the storage cluster and delivers exabyte scale at 80+ GB/s to each GPU in the client layer.

Automated tiering expands the Hitachi Content Software for File namespace from fast flash to economical hard disk storage via on-premises or hybrid cloud object storage to optimize scalability and cost. This distributed design also eliminates the bottlenecks of traditional data protection.

Data protection and business continuity are designed into the solution with logical snapshots, data protection levels, erasure coding and remote replication. HCP brings built-in ransomware protection in the form of immutability, versioning, and encryption. HCP also addresses data governance requirements with robust retention, destruction, authenticity, access controls, and auditing capabilities.

The following illustration shows Hitachi Content Software for File integration with Hitachi Content Platform.



Hitachi iQ compute nodes

Hitachi iQ compute nodes include Supermicro GPU SuperServer SYS-821GE-TNHR systems that are a NVIDIA certified GPU-accelerated system for all AI and HPC workloads, offering unprecedented compute density, performance, and flexibility. They include 8 × NVIDIA H100 SXM GPUs.



Key Features:

- Up to 8 onboard NVIDIA SXM: HGX H100 Tensor Core GPUs (80 GB) or HGX H200 GPU (141 GB)
- 5th/4th Gen Intel® Xeon® Scalable processor support
- 32 DIMM slots up to 8TB: 32 × 256 GB DRAM Memory Type: 5600MTs ECC DDR5
- 8 PCIe Gen 5.0 X16 LP
- 2 PCIe Gen 5.0 X16 FHHL Slots (optional)
- Flexible networking options
- 2 M.2 NVMe for boot drive only
 - 16 × 2.5" Hot-swap NVMe drive bays (12 by default, 4 optional)
 - 3 × 2.5" Hot-swap SATA drive bays
 - Optional: 8 × 2.5" Hot-swap SATA drive bays
- 10 heavy duty fans with optimal fan speed control
- Optional: 8 × 3000W (4+4) Redundant Power Supplies, Titanium Level
- 6 × 3000W (4+2) Redundant Power Supplies, Titanium Level

The rear ports of HGX H100 system are described in [Compute node \(HGX H100\) network overview \(on page 18\)](#).

See the [Supermicro documentation](#) for details.

Hitachi iQ networking

NVIDIA Quantum-2 QM9700 switch

NVIDIA Quantum-2 QM9700 switches with 400 Gb/s NDR InfiniBand connectivity are used for setting up compute and storage fabric in this solution. NVIDIA ConnectX-7 single-port adapters are used for connecting HGX compute nodes and HCSF storage nodes to InfiniBand fabric at 400 Gb/s speed.



NVIDIA Spectrum-3 SN4600 switch

NVIDIA Spectrum-3 SN4600 Ethernet switches are used for setting up the in-band management network. It offers 64 ports at speeds between 1 Gb/s and 200 Gb/s.



NVIDIA Spectrum SN2201 switch

NVIDIA Spectrum SN2201 Ethernet switches are used for setting up the Out-Of-Band management network. They offer 48 ports at 1 Gb/s speed.



NVIDIA ConnectX-7[®] Adapters

ConnectX-7 Adapters can provide 25/50/100/200/400 Gb/s of throughput.

The HGX H100 compute host and HCSF storage nodes are equipped with single port ConnectX-7 InfiniBand HCA to provide 400 Gb/s throughput for compute and storage fabric.

NVIDIA ConnectX-6[®] NICs

ConnectX-6 NICs can provide 10/25/40/50/100/200 Gb/s of throughput.

The HGX H100 compute host and management host use dual port ConnectX-6 100 Gb/s NICs for high speed in-band management network connectivity.

NVIDIA LinkX[®] cables

The NVIDIA LinkX product family of cables and transceivers provides the industry's most complete line of 10, 25, 40, 50, 100, 200, and 400 Gb/s in Ethernet and 100, 200 and 400 Gb/s InfiniBand connectivity options.

Kubernetes

Kubernetes is an open-source container orchestration platform for deployment automation, scaling, and management of containerized applications.

NVIDIA GPU Operator

The NVIDIA GPU Operator uses the operator framework within Kubernetes to automate the management of all NVIDIA software components needed to provision GPUs. These components include the NVIDIA drivers (to enable CUDA), Kubernetes device plugins for GPUs, the NVIDIA Container Runtime, automatic node labeling, DCGM-based monitoring, and more.

NVIDIA Network Operator

An analog to the NVIDIA GPU Operator, the NVIDIA Network Operator simplifies scale-out network design for Kubernetes by automating aspects of network deployment and configuration that would otherwise require manual work. It loads the required drivers, libraries, device plugins, and CNIs on any cluster node with an NVIDIA network interface. Paired with the NVIDIA GPU Operator, the Network Operator enables NVIDIA Magnum IO GPUDirect RDMA, a key technology that accelerates cloud-native AI workloads by orders of magnitude. The NVIDIA Network Operator uses Kubernetes CRD and the Operator Framework to provision the host software needed for enabling accelerated networking.

NVIDIA AI Enterprise

NVIDIA AI Enterprise is a comprehensive, cloud-native software platform designed to accelerate data science pipelines and streamline the development and deployment of production-grade co-pilots and generative AI applications. It offers easy-to-use microservices that deliver optimized model performance, coupled with enterprise-grade security, support, and stability. This ensures a seamless transition from prototype to production for businesses that rely on AI to operate.

NVIDIA Base Command Manager Essentials

NVIDIA Base Command™ Manager Essentials, a part of the NVIDIA AI Enterprise software portfolio, provides rapid deployment and comprehensive management for diverse AI and high-performance computing (HPC) clusters, whether at the edge, in data centers, or in multi- and hybrid-cloud environments.

Base Command Manager Essentials automates the provisioning and administration of clusters from a few nodes to hundreds of thousands, supports systems with NVIDIA GPU acceleration and other configurations, and enables orchestration through Kubernetes.

RDMA

Remote Direct Memory Access (RDMA) is a technology that allows computers in a network to exchange data without involving the processor, cache or operating system of either computer.

Like locally based DMA, RDMA improves throughput and performance and frees up compute resources.

NVIDIA Magnum IO™ GPUDirect® RDMA

NVIDIA Magnum IO GPUDirect (GDR) RDMA provides a direct P2P (Peer-to-Peer) data path between the GPU memory directly to and from NVIDIA ConnectX adapters. This reduces GPU-to-GPU communication latency and completely offloads the CPU, removing it from all GPU-to-GPU communications across the network.

NVIDIA Magnum IO GPUDirect Storage

NVIDIA Magnum IO GPUDirect Storage (GDS) provides a direct memory access (DMA) path between the memory of two graphics processing units (GPUs) and GPUDirect RDMA, enabling a direct DMA path to a network interface card (NIC). GDS creates a direct DMA data path between GPU memory and storage, thus avoiding a bounce buffer through the CPU.

This direct path can increase system bandwidth while decreasing latency and utilization load on the CPU.

GPUDirect Storage offers the following capabilities:

Direct Memory Access (DMA) Engine: DMA capabilities allow for direct communication between GPU memory and storage devices, bypassing the need to copy data through system memory. This reduces latency and enhances overall system performance.

RDMA Capabilities: GPUDirect Storage utilizes RDMA technology to efficiently access data stored in remote memory locations without involving the CPU, enabling data transfer between GPUs and storage devices across the network.

NVIDIA Kernel Extensions and Drivers: These extensions and drivers facilitate the integration of GPUDirect Storage, enabling efficient data transfer paths between storage and GPU memory.

Coherent Memory Access: GPUDirect Storage ensures consistent memory access and data integrity between GPUs and storage devices during data transfers.

Solution components

These are the key hardware and software components used during testing. For the latest configuration and version details, consult your Hitachi Vantara technical representative to verify the most current information.

Hardware components

The following table lists the hardware and OS/firmware versions that were tested.

Vendor	Hardware	Detail Description	Version	Quantity
Hitachi Vantara	HA810 G3	Management servers: 2 × Head nodes 3 × K8s Master nodes	iLO FW: 1.53 BIOS: 1.48 OS: RHEL 9.2	5
		CPU: 2 × Intel® Xeon® Gold 5418Y (185W, 24C,2.0GHz) Total Cores 48, Threads 96		
		RAM: 16 × 32 GB DIMMs (512 GB)		
		Drive: <ul style="list-style-type: none"> ▪ Internal SSD - 1 × 6.4 TB NVMe (optional) ▪ M.2 boot device - 2 × 480 GB NVMe SSD (NS204i-u Boot controller) 		
		NIC: <ul style="list-style-type: none"> ▪ 2 × 100 Gb 2p CX-6 cards ▪ 1 × 10/25GbE 2p SFP28 		
Super Micro Computer	SYS-821GE-TNHR	GPU SuperServer SYS-821GE-TNHR	BMC FW:01.01.04 BIOS:2.1 OS: RHEL 9.2	2
		CPU: 2 × Intel® Xeon® Platinum 8480+ Processor 105M Cache, 2.00 GHz, Total Cores 56, Total Threads 112		
		GPU: 8 × NVIDIA H100 SXM 80 GB Tensor Core		
		GPU memory: 640 GB		
		Memory: 32 × 64 GB DIMMs (2 TB)		

Vendor	Hardware	Detail Description	Version	Quantity
		Drive: <ul style="list-style-type: none"> ▪ M.2 Boot device - 2 × Micron 7450 PRO 1.9 TB NVMe PCIe 4.0 M.2 22 × 110mm 3D TLC ▪ Internal SSD - 2 × Kioxia CD6-R 1.92 TB NVMe PCIe 4×4 2.5" 15mm SIE 1DWPD ▪ Internal SSD - 8 × Kioxia CD6-R 3.84 TB NVMe PCIe 4×4 2.5" 15mm SIE 1DWPD 		
		NIC: <ul style="list-style-type: none"> ▪ 8 × NVIDIA 900-9X766-003N-SQ0 PCIe 1-port IB and Ethernet 400 GbE OSFP Gen5 × 16 CX7, RoHS ▪ 1 × PCIe 2-port 100 Gb/s InfiniBand or Ethernet, QSFP56, Gen 3.0/4.0 x16, CX-6 VPI ▪ 2 × NVIDIA 900-9X766-003N-SQ0 PCIe 1-port IB and Ethernet 400 GbE OSFP Gen5x16 CX7,RoHS ▪ 1 × standard low-profile 2-port 10 GbE RJ45 based on Intel X550-AT2 		
Super Micro Computer	SMC Gen 5 Storage A+ Server ASG-1115S-NE316R (31116)	HCSF 36116 Storage Nodes CPU: AMD EPYC 9534 64-Core Processor 2.45G 256M 280W SP5 RAM: 768 GB 12 × 64 GB DDR5 - 4800 Drive: <ul style="list-style-type: none"> ▪ 15 × 3.84 TB NVMe ▪ 1 × 960 GB NVMe for boot 	BMC FW: 01.02.23 BIOS: 1.6b OS: Rocky 8.8	12

Vendor	Hardware	Detail Description	Version	Quantity
		NIC: <ul style="list-style-type: none"> ▪ 2 × CX7 IB 400 Gb OSFP ▪ AIOM Dual Port 10G ▪ AIOM Dual Port 25G 		
NVIDIA	NVIDIA QM9700	64 port 400 Gb/s InfiniBand switches for Compute and Storage Fabric 64-ports NDR, 32 OSFP ports	MLNX OS 3.11	4
NVIDIA	NVIDIA SN4600	64 port 200 Gb/s Ethernet switch for In-Band Management 64-port 200 GbE QSFP56 Splittable to up to 128 X 10/25/50/100 GbE ports when used with splitter cables	Cumulus Linux 5.3.1	2
NVIDIA	NVIDIA SN2201	48 port 1 Gb/s Ethernet switch for OOB management	Cumulus Linux 5.2.0	1

Software components

The following table lists the key software components that were tested.

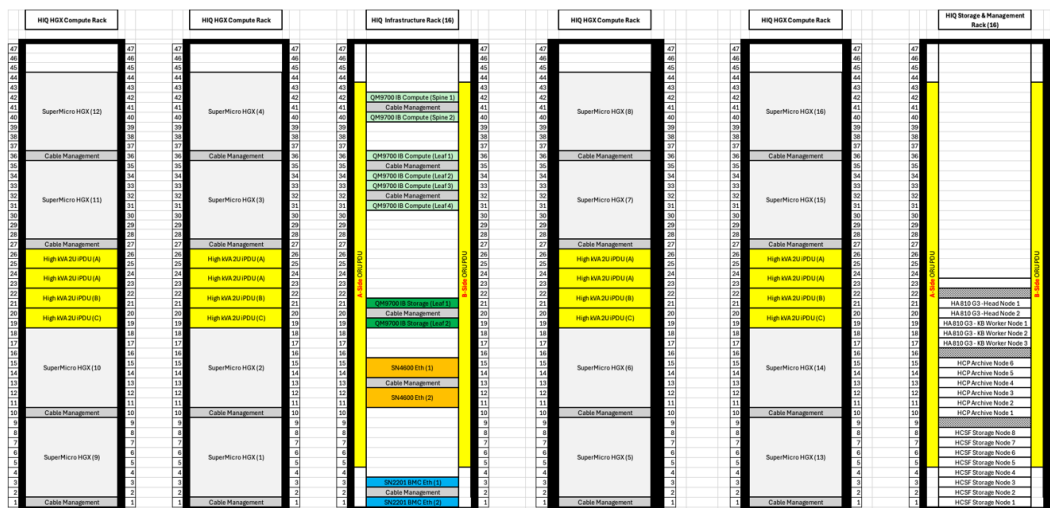
Component	Version
NVIDIA Base Command Manager Essentials version	10 (RHEL 9.2)
OS	Management Nodes: RHEL 9.2 HGX Nodes: RHEL 9.2 HCSF Node: Rocky 8.8
Kubernetes and Operator	Upstream Kubernetes (K8s) 1.28 NVIDIA GPU Operator 23.9.2 NVIDIA Network Operator 23.7.0
NVIDIA Drivers and Tools	Linux Kernel: 5.14.0-284.30.1 NVIDIA Open Driver Version: 550.90.07 CUDA Version: 12.4 GDS release version: 1.10.1.7

Component	Version
	NVIDIA_fs version: 2.20 libcufile version: 2.12
OFED	MLNX_OFED_LINUX-23.10-0.5.5.0
CX7 Driver	28.39.2048
HCSF Software	v4.2.10.30-hcsf
Spectrum Ethernet Switch	NVIDIA Cumulus Linux 5.3.1
Quantum IB Switch	MLNX OS 3.11

Solution design

This is a detailed solution example of how Hitachi iQ enterprise solution with NVIDIA HGX H100 and Hitachi Content Software for File (HCSF) storage is configured.

The following is a reference rack elevation design for 16 HGX compute nodes. This reference design offers a scalable and modular architecture to start as small as 2 nodes and grow up to 16 nodes. The rack layout can be adjusted to meet local data center requirements, such as maximum power per rack and rack layout between system, storage, and network components to meet local needs for power and cooling distribution.

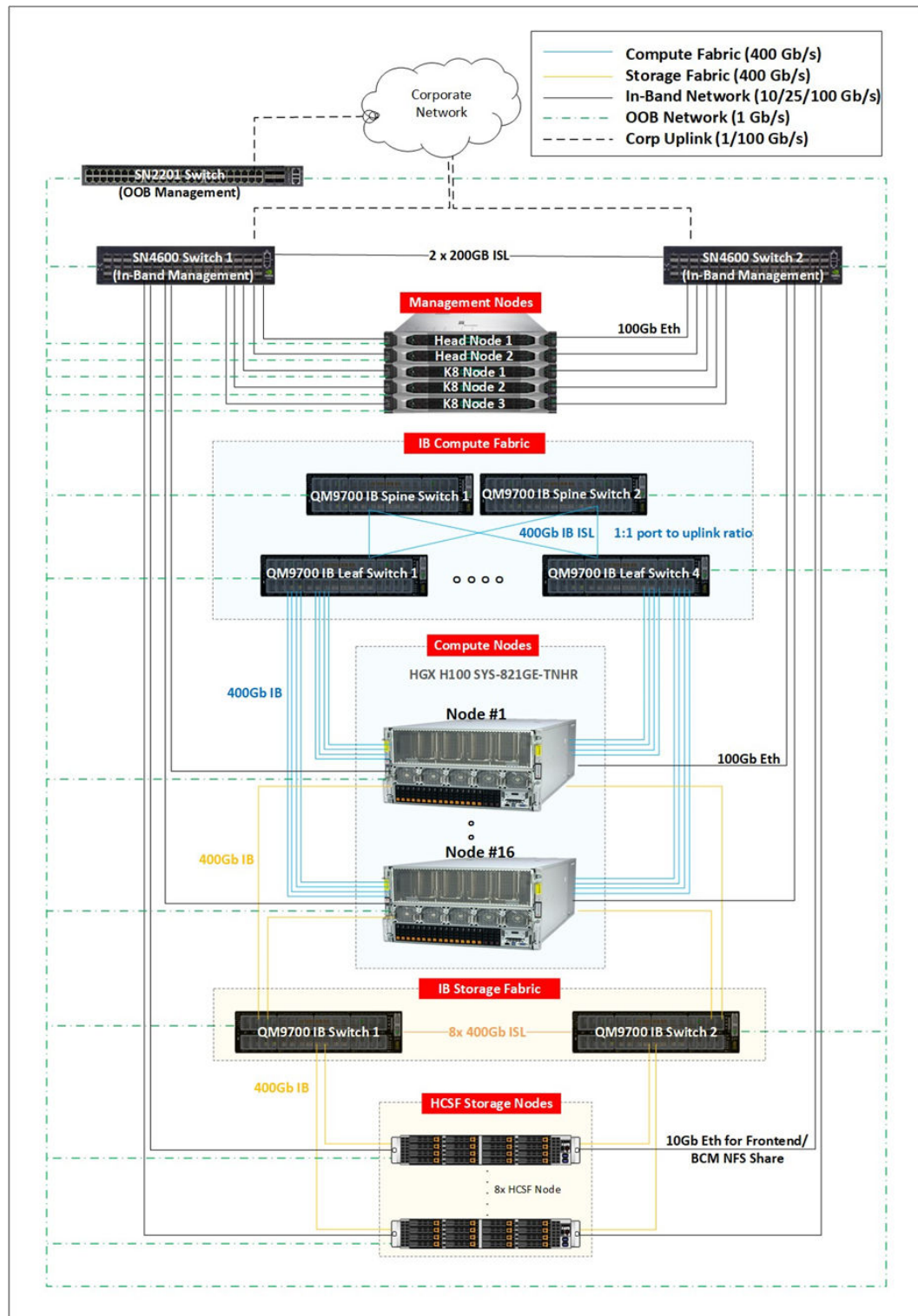


Physical architecture

This section provides an abstract description of how 16 × GPU SuperServer SYS-821GE-TNHR with HCSF and other management servers are connected and configured.

- Fully redundant HW and network connectivity (HA setup)
- High-performance NVIDIA Quantum-2 400 Gb/s InfiniBand for compute and storage fabric

- 100 Gb/s redundant Ethernet connectivity for in-band management network
- Out-of-band (OOB) management connectivity with 1 Gb/s Ethernet
- Optional Hitachi Content Platform (HCP) for object storage capabilities, data tiering, and data protection



In this architecture, the network comprises NVIDIA Quantum-2 QM9700 400 Gb InfiniBand (IB) switches for storage and computing fabric setup. The HCSF nodes are directly connected to the QM9700 IB switches (storage fabric) for faster connectivity from the HGX to the storage layer to leverage the GDS feature of the HCSF node.

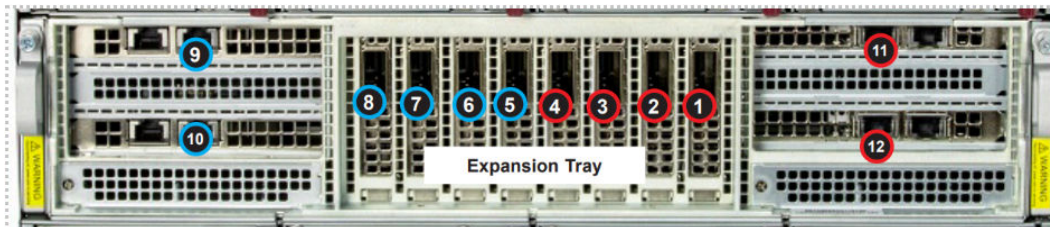
Compute node (HGX H100) network overview

On a high level, four different networks are typically configured.

- Compute Fabric: InfiniBand network using NVIDIA Quantum-2 QM9700 switch for connecting all HGX servers for GPU to GPU connectivity.
- Storage Fabric: InfiniBand network using NVIDIA Quantum-2 QM9700 switch for connecting HGX servers and HCSF storage.
- In-Band Management Network: Network used exclusively within the cluster, for in-band management using NVIDIA Spectrum-3 SN4600 switch.
- Out-of-Band Management Network: Network for out of band management, connecting BMC/iLO ports on NVIDIA Spectrum SN2201 switch.

Optionally, an external network can be configured to uplink the stack to corporate network and internet access for more secure and private deployments.

These are the available ports on the HGX server expansion tray that will be used for compute/storage fabric and in-band management network as described in the following table.

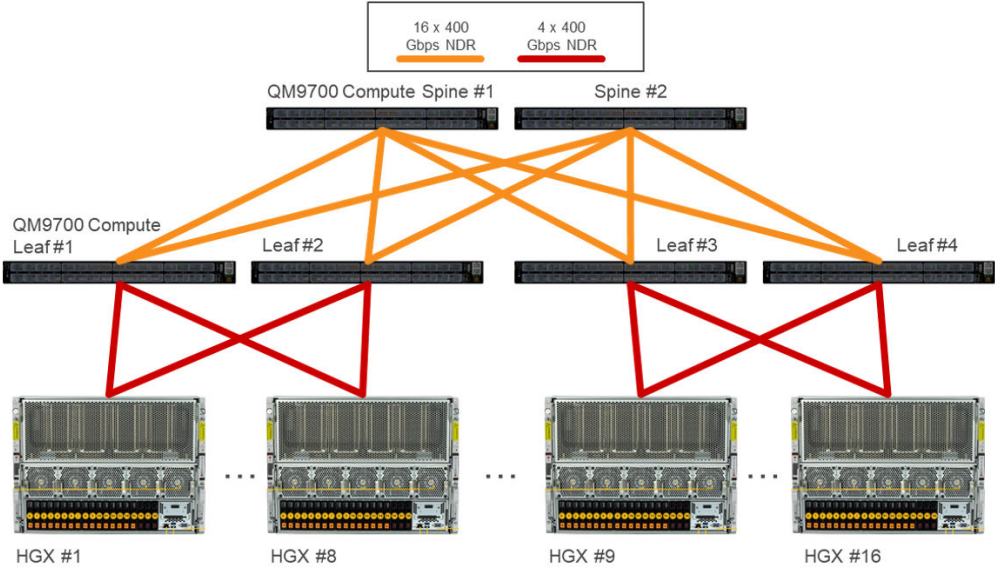


Slot Number	NIC Adapter	Purpose
1	ConnectX-7 400G NDR InfiniBand	Compute Fabric Switch 1
2	ConnectX-7 400G NDR InfiniBand	Compute Fabric Switch 1
3	ConnectX-7 400G NDR InfiniBand	Compute Fabric Switch 1
4	ConnectX-7 400G NDR InfiniBand	Compute Fabric Switch 1
5	ConnectX-7 400G NDR InfiniBand	Compute Fabric Switch 2
6	ConnectX-7 400G NDR InfiniBand	Compute Fabric Switch 2
7	ConnectX-7 400G NDR InfiniBand	Compute Fabric Switch 2
8	ConnectX-7 400G NDR InfiniBand	Compute Fabric Switch 2
9	ConnectX-6 100G Ethernet/InfiniBand Dual Port	High speed In-band Management

Slot Number	NIC Adapter	Purpose
10	ConnectX-7 400G NDR InfiniBand	Storage Fabric Switch 2
11	Intel X550 Dual Port 10G	Optional
12	ConnectX-7 400G NDR InfiniBand	Storage Fabric Switch 1

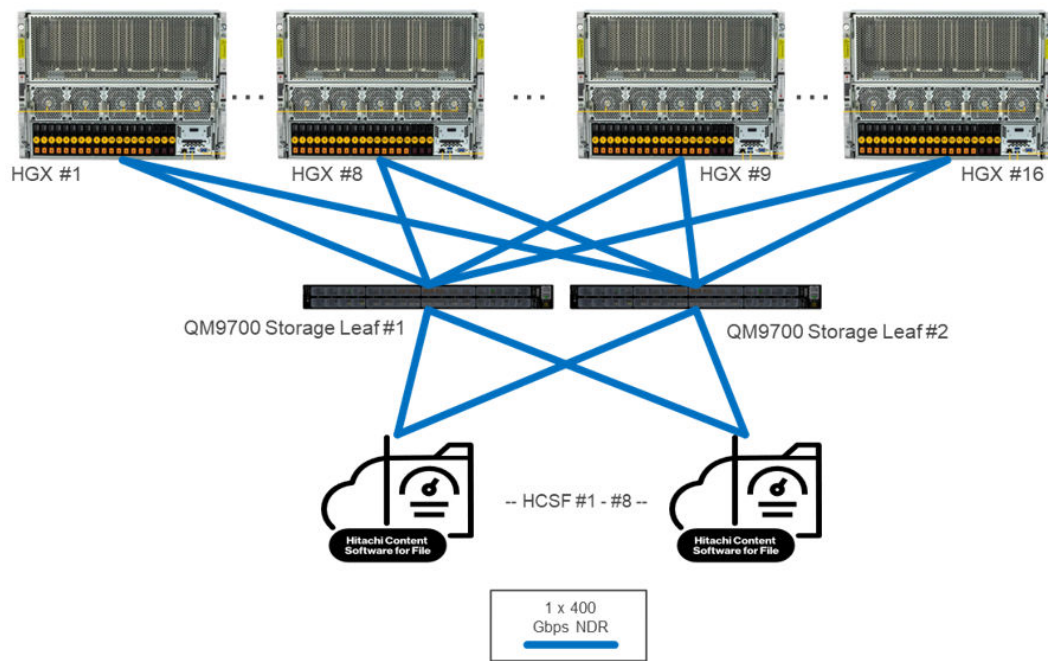
InfiniBand compute fabric

Compute fabric is built for GPU-to-GPU communication using full flat-tree topology using NVIDIA Quantum-2 QM9700 400Gb/s (NDR) InfiniBand switches. See the following reference design for 16 HGX nodes that are connected using 4 leaf switches and 2 spine switches. Slot number 1 to 8 on HGX servers are used for compute fabric connectivity.



InfiniBand storage fabric

The storage fabric provides high bandwidth to shared HCSF storage using NVIDIA Quantum-2 400Gb/s InfiniBand QM9700 switches. Each storage and compute node uses 2 x 400 Gb IB links with IPoB for communication. The following diagram illustrates connectivity for 16 HGX nodes and 8 HCSF storage nodes. Slot number 10 and 12 on HGX servers are used for storage fabric connectivity.

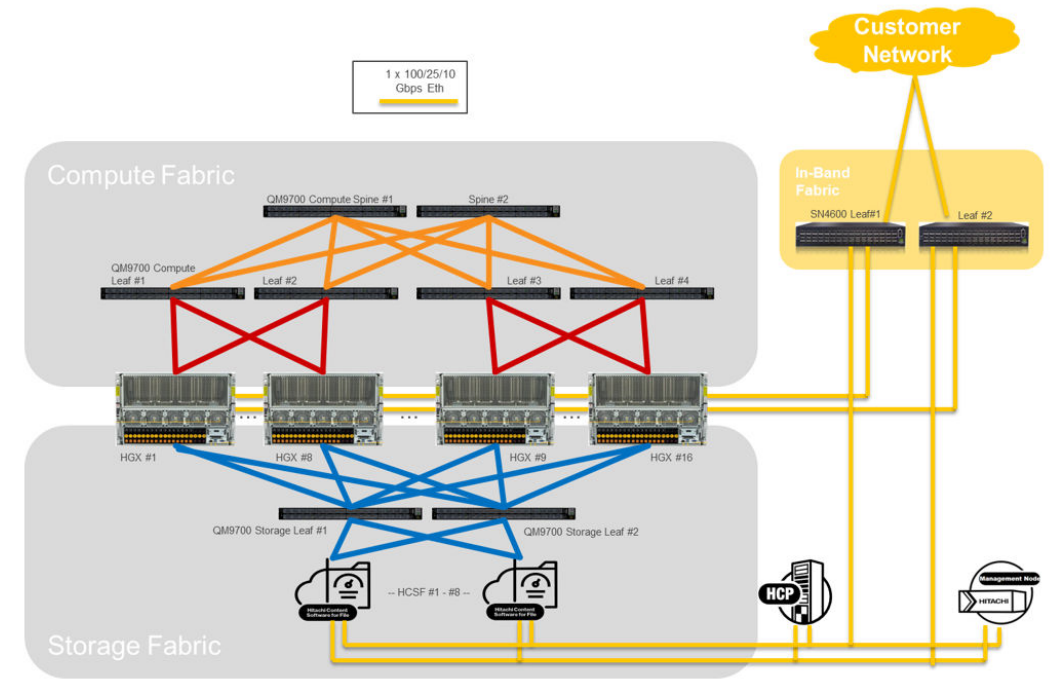


In-band management network

The in-band network is configured using NVIDIA Spectrum-3 SN4600 switches. Compute and management nodes use a 100 Gb/s high speed in-band Ethernet network for inter-node communication and cluster management. Slot number 9 on the HGX server with a 2-port 100 Gb/s card is connected to an in-band as bonded interface for high availability. Similarly, the K8s control plane nodes and the head nodes have a bonded interface configuration connected to SN4600 switches.

HCSF storage nodes are connected using a 10 Gb/s link on the in-band network for management and NFS export setup. An NFS share is mounted on management head nodes over the in-band network for high availability setup of NVIDIA Base Command Manager software.

Optionally HCSF Storage Nodes also connect to the HCP platform using a 25 Gb/s link over the in-band network for object storage capabilities, data tiering, and data protection.

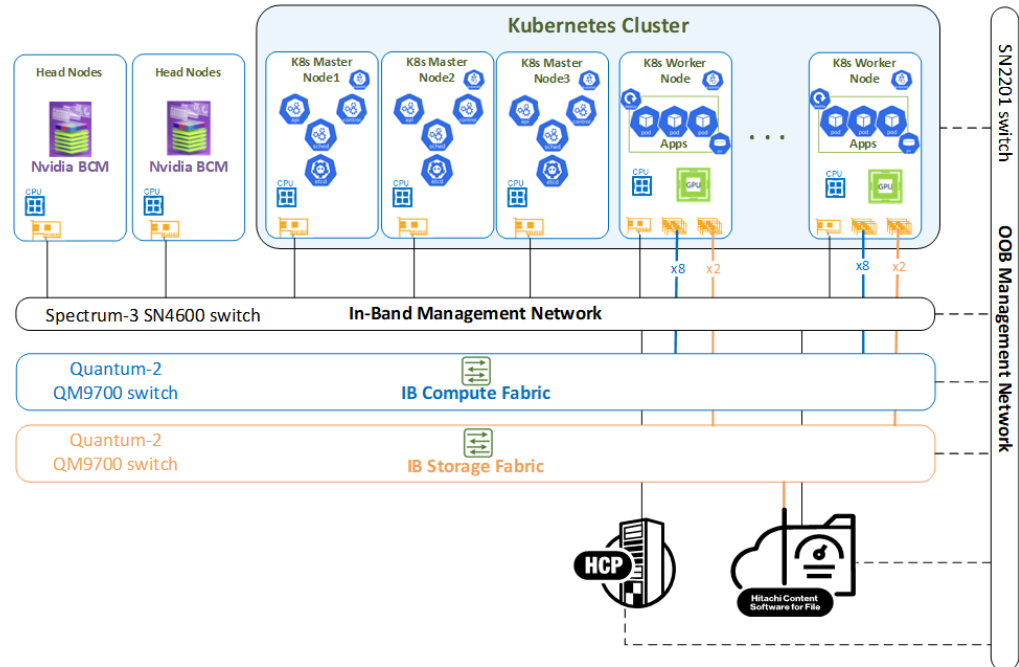


Out-of-band management network

OOB Ethernet fabric connects the management ports of all devices including HGX, management servers, storage, networking gear, rack PDUs, and all other devices. These are separated onto their own fabric because there is no use-case where users need access to these ports and are secured using logical network separation for remote access management. OOB network is set up using NVIDIA Spectrum SN2201 1 Gb/s Ethernet network switches.

Logical design

The following illustration shows the Hitachi iQ HGX solution logical design for cloud native deployments using the NVIDIA BCM software.



The logical design includes the following:

- 2 × head/deployment nodes running NVIDIA Base Command Manager Essentials in HA mode on Hitachi Advanced Server HA820 G3 servers, that helps provision other nodes and deploy Kubernetes clusters.
- A high availability cluster consisting of multiple control plane nodes (K8s master nodes) and multiple worker nodes (HGX H100 servers).
 - 3 × K8s primary nodes running all Kubernetes management components on HA820 G3 servers.
 - 2 to 16 × HGX H100 servers K8s worker nodes.
- InfiniBand fabric for compute and storage connectivity using NVIDIA Quantum-2 QM9700 switches.
- In-band management network for deployment and K8s management networks using an NVIDIA Spectrum-3 SN4600 switch.
- Out-of-band management network for BMC/iLO remote access using an NVIDIA Spectrum SN2201 switch.
- Hitachi Content Software for File (HCSF) storage.
- Optional Hitachi Content Platform (HCP) storage

Solution verification

Hitachi iQ enterprise HGX solution was validated using 2 HGX H100 compute nodes according to the [Solution design \(on page 16\)](#) using InfiniBand network for compute and storage fabric.

As part of solution verification, the following key points were verified:

- Using NVIDIA Base Command Manager (BCM) Essential to accelerate deployment and management of the solution infrastructure.
- Solution scalability and high availability testing.
- MLperf inference performance benchmarking to test GPU-based workloads.
- GPU Direct Storage (GDS) configuration and its performance benchmarking using the GDSIO utility.
- Storage benchmarking for Hitachi Content Software for File (HCSF) using the FIO utility.

Hitachi Content Software for File Storage also got BasePOD certification during this period from NVIDIA. See [Hitachi Achieves NVIDIA DGX BasePOD™ Certification](#) for details.

The following table summarizes the key lab verification tests performed for HCSF storage, which exceed performance expectations.

Performance Characteristic	NVIDIA Recommendation (GB/s)	Hitachi Results (GB/s)
Hitachi Performance Comparison		
Single node read	40	60
Single node write	20	60

Hitachi Content Software for File Performance	
Workload Type	Peak Performance
Sequential read performance	745 GB/s
Sequential write performance	257 GB/s
Random 4kB read performance	26.2 MIOPS
Random 4kB write performance	6.16 MIOPS
4kB read latency	112μs
4kB write latency	78μs

Conclusion

Hitachi iQ enterprise solution offers unparalleled computational power of NVIDIA HGX systems with the scalable and efficient storage capabilities of Hitachi Content Software for File offering a comprehensive solution that addresses the complexities of modern high-performance computing and AI-driven workloads.

This reference architecture serves as a foundational guide, providing the necessary tools and best practices to leverage the combined strengths of NVIDIA HGX systems and Hitachi Content Software for File Storage, ultimately driving success in the era of data-driven decision-making.

Key benefits of this reference architecture include:

- Enhanced performance and scalability for demanding AI and high-performance computing (HPC) applications.
- Reliable and secure data management, ensuring data integrity and compliance.
- Simplified deployment and management, reducing operational complexities.
- Comprehensive support for advanced AI workflows, driving innovation and insights.

By adopting this solution, organizations can effectively harness the power of AI and HPC, transforming data into actionable insights and achieving strategic objectives with greater efficiency.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA



HitachiVantara.com/contact