Hitachi Vantara

Datasheet

Hitachi iQ with NVIDIA HGX

Hitachi's AI Solutions Stack for Enterprise Workloads

Hitachi iQ with NVIDIA HGX™ integrated solutions delivers extreme performance and resiliency at scale, making your AI infrastructure simpler and faster to design, deploy, and manage.

Organizations are constantly looking for ways to automate, accelerate time-to-market, and develop new insights, products, or innovations to propel their business forward. At the same time, AI and Generative AI technologies are revolutionizing industries by enhancing existing capabilities and transforming how quickly and creatively business problems can be solved. Whether building specific AI solutions or just starting to identify general-purpose capabilities, Hitachi iQ has the power to automate your business processes and improve your AI experience.

Hitachi iQ allows organizations to automate, expedite, and streamline their business through intelligent, performant, scalable and flexible AI and GenAI solutions. Unlike conventional AI offerings, 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 each organization's unique needs and objectives.

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. They optimize AI deployments with end-to-end software-defined AI and data analytics software to streamline the development and deployment of production-grade AI applications from pilot to production.

Hitachi iQ with NVIDIA HGXTM configurations can be purchased in whole from Hitachi and does not require engaging other vendors to build the AI solutions stack.

- **Personalize Your AI**
  Built to meet the rigors of AI and engineered for industry outcomes but customized to your needs.
- **Achieve Faster Insight**
  Accelerated architecture delivers storage performance that improves GPU resource utilization.
- **Simplify to Scale**
  Validated reference architecture blueprints provide the flexibility and scale to rapidly develop, test, and deploy modern AI solutions.
- **Lower TCO**
  Leverage lower cost, erasure coded, scale-out object storage to safeguard data, catalog for reuse, and store data long term.
- **Improve Accuracy**
  Increase the quantity and quality of data to improve the reliability of results. Identify, classify, transform, move, consolidate and prepare data to get the most value out of AI/ML

Fast and Efficient Results

Harness the power of validated and seamlessly integrated industry-leading AI technologies to accelerate your transformation journey and quickly gain insights.

Timely and Accurate Decisions

Ensure data is relevant and accessible for AI applications and analytics while prioritizing security and compliance.

Industry Relevant, Meaningful Outcomes

Recognizing that not all organizations possess the AI skills required for their industry, we provide the expertise to help bridge the gap, creating cutting-edge AI applications and improving your outcomes.

Hitachi iQ is the Flagship AI Portfolio from Hitachi, Powered by NVIDIA® Technologies.

When customers try to implement their own AI solutions, they are often faced with the challenges of creating complicated systems and performing the necessary integrations on their own. Instead, Hitachi iQ with NVIDIA HGX offers pre-built platforms with bespoke customization and delivery—all from a single vendor. Engineered with Hitachi and NVIDIA solutions, Hitachi iQ is an accelerated integration based on proven foundations. For AI accelerated compute, Hitachi iQ with NVIDIA HG platforms combine outstanding software, infrastructure, and expertise in unified and scalable AI development solutions. For primary storage, Hitachi Content Software for File (HCSF) provides a distributed parallel file system that delivers the highest performance file services by leveraging NVMe flash and also includes integrated tiering that seamlessly expands a single namespace to and from hard disk drive (HDD) object storage without the need for special data migration software or complex scripts. Object storage, data tiering, and data protection can be optionally provided by the Hitachi Content Platform (HCP).

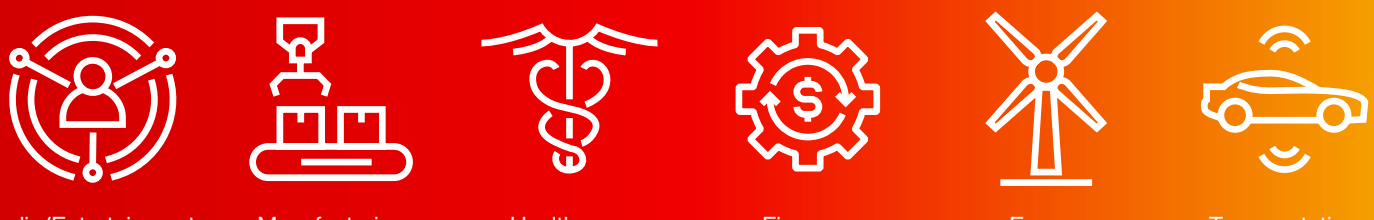
Hitachi iQ solutions also include NVIDIA-specified compute and storage infrastructure and interconnect components, which can be completed with the NVIDIA Base CommandTM Manager software suite for provisioning, managing, and monitoring and NVIDIA AI Enterprise for AI deployment and enablement. Hitachi provides end-to-end services and the support of our partner ecosystem.

Workloads and Use Cases

Whether you are looking for industry-specific AI solutions or just starting to identify general-purpose capabilities, AI has the power to automate your business processes and improve your customer experience. The industry already uses AI to create art and graphic design elements, write code and generate marketing tag lines. And new use cases are being identified daily that can provide immediate benefits to any organization. For example:

- Customer Service Voice Assistant: Revolutionizes customer service with its advanced voice recognition technology, offering real-time responses and personalized assistance.
- Large Language Model (LLM) Recommender System: Delivers highly accurate, context-aware suggestions and personalizes content, products, and services, making discovery seamless and engaging.
- Coding and Development Copilot: Can act as an invaluable partner in the development process, offering real-time suggestions, debugging assistance, and code optimization, accelerating development cycles and enhancing code quality.
- Automated Document Processing and Analysis: Streamlines the processing, analysis, and management of large volumes of documents within enterprise environments, significantly reducing manual workloads, improving accuracy and enhancing decision-making.
- Financial Reporting and Accounting: Can reduce the repetitive tasks that workers and consultants are required to do, helping streamline operations and reduce errors for data entry, transaction categorization and invoice processing.
- Edge Inference: Brings AI computing closer to the data source, minimizing latency and enhancing real-time decision-making. It's ideal for applications requiring instant analysis and action, from autonomous vehicles to smart city infrastructure.

Hitachi iQ leverages its deep domain experience with 110+ years of experience in the Operational Technology (OT) market in finance, transportation, energy, media, entertainment, manufacturing and healthcare. With Hitachi's OT and IT legacy, the Hitachi iQ portfolio integrates infrastructure and capabilities to deliver industry-specific AI solutions tailored to our customers' needs.



Hitachi iQ Solution Architecture

Hitachi iQ includes the testing and qualification of NVIDIA HGX systems with Hitachi Content Software for File primary storage, creating a robust solution for high-performance computing (HPC) and advanced AI workloads. Hitachi testing of the full Hitachi iQ with NVIDIA HGX reference architecture validates the seamless integration and interoperability between NVIDIA's AI hardware and Hitachi's scalable file storage system, ensuring optimal performance, reliability and supportability.

For more information and details about the Hitachi iQ reference architecture for NVIDIA HGX systems, please refer to "Hitachi iQ with NVIDIA HGX H100 and Hitachi Content Software for File Storage" on the Hitachi Vantara website.

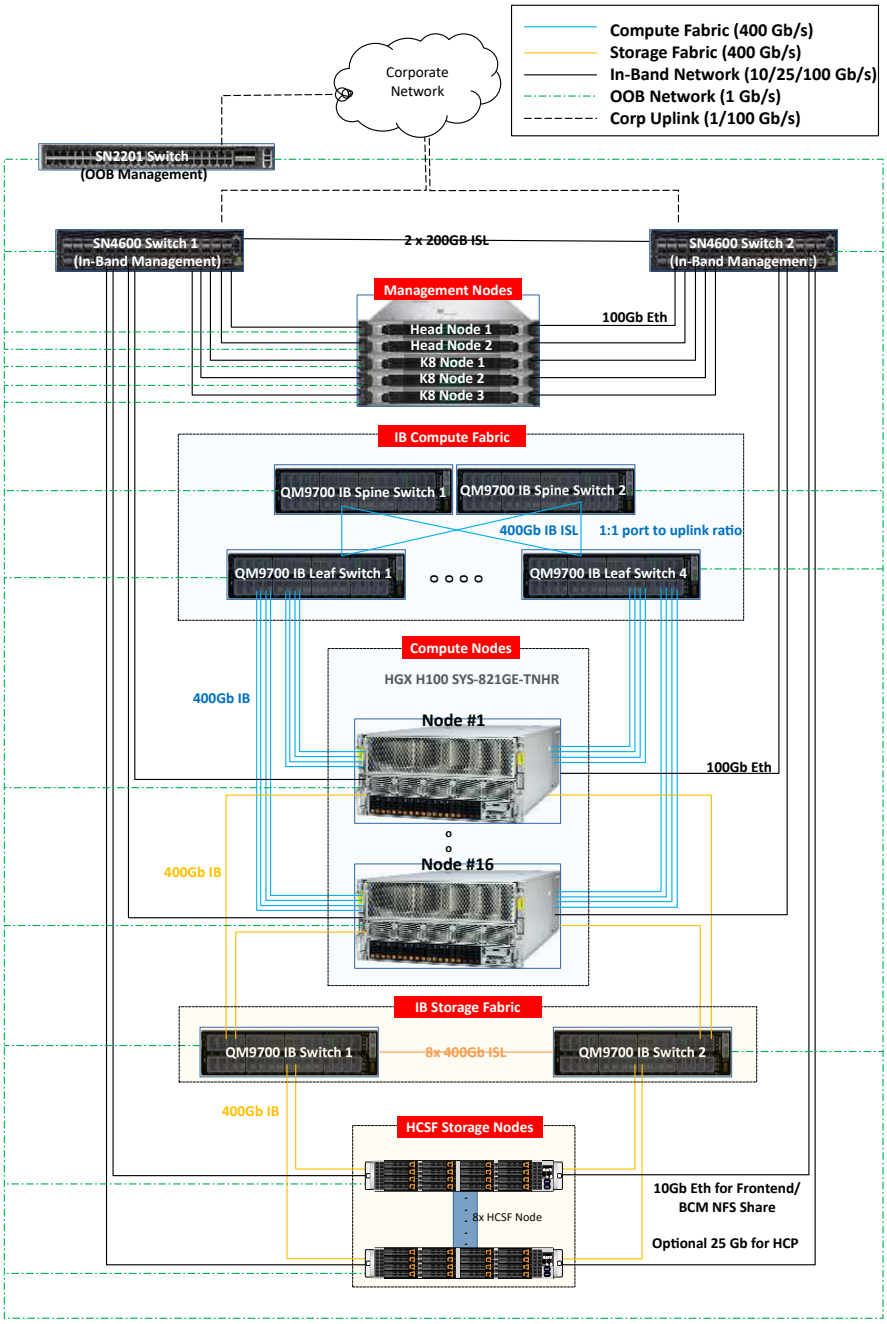


Diagram 1: Hitachi iQ with NVIDIA HGX Reference Architecture

NVIDIA Architecture Components	**NVIDIA HGX H100 System:** Supermicro GPU SuperServer SYS-821GE-TNHR powered by eight NVIDIA H100 Tensor Core GPUs per node in an 8U chassis with advanced thermal capacity. It can scale up to 128 HGX nodes.
NVIDIA Software	**NVIDIA Base Command™ Manager:** Provides rapid deployment and comprehensive management for diverse AI and high-performance computing (HPC) clusters. **NVIDIA AI Enterprise:** Comprehensive, cloud-native software platform designed to accelerate data science pipelines and streamline the development and deployment of production-grade copilots and generative AI applications. **NVIDIA GPU Direct Storage:** Enables memory access and coherency features that bypass system CPUs and/or memory and ensures data consistency and integrity between GPUs and storage devices.
Primary Storage	**Hitachi Content Software for File 36116 Storage Nodes:** Hypervisor-based, software-defined storage overcomes traditional storage scaling and file sharing limitations but also allows unified file access via POSIX, NFS, SMB, S3 and NVIDIA® GPUDirect® Storage.
Networking	**NVIDIA QM9700 400Gb/s InfiniBand Switches**: Compute and storage fabric. **NVIDIA SN4600 200GbE Ethernet Switches:** Storage fabric for management servers and in-band management network. **NVIDIA SN2201 1GbE Ethernet Switch:** Out-of-band management network.
Management servers	**Hitachi Vantara HA810 G3 Servers**

Table 1: Hitachi iQ with NVIDIA HGX Components

Hitachi iQ Solution Performance*	Performance Characteristic	NVIDIA "Best" Recommendation		Hitachi iQ Performance		
	Single-node read Single-node write	40 GB/s 20 GB/s		60 GB/s 60 GB/s		
Hitachi Content Software for File Performance*	Workload Type	Peak Performance				
	Sequential read performance Sequential write performance Random 4kB read performance Random 4kB write performance 4kB read latency 4kB write latency	745 GB/s 257 GB/s 26.2 MIOPS 6.16 MIOPS 112µs 78µs				
MLPerf v4.0 Benchmark Scores (Lab testing results, not published)	bert-99 llama2-70b-99.9 Stable-diffusion-xl	70831.7 Samples/s 21206.1 Tokens/s 13.29 Samples/s				
NVIDIA GPUDirect Storage (GDS) Performance (Measured with one HGX server using NVIDIA gdsio utility)		Sequential read			Sequential write		
	I/O size	CPU_GPU throughput	GDS-enabled throughput	GDS boost	CPU_GPU throughput	GDS-enabled throughput	GDS boost
	32 KiB	9.1 GiB/s	9.9 GiB/s	8.2%	16.2 GiB/s	33.4 GiB/s	105.8%
	64 KiB	15.3 GiB/s	19.7 GiB/s	28.6%	22.7 GiB/s	58.8 GiB/s	159.3%
	128 KiB	24.8 GiB/s	38.1 GiB/s	53.8%	30.4 GiB/s	73.8 GiB/s	143.2%
	256 KiB	36.3 GiB/s	47.3 GiB/s	30.1%	55.2 GiB/s	74.7 GiB/s	35.2%
	512 KiB	47.2 GiB/s	53.6 GiB/s	13.5%	67.5 GiB/s	75.0 GiB/s	11.1%
	1024 KiB	52.8 GiB/s	64.6 GiB/s	22.3%	66.8 GiB/s	71.2 GiB/s	6.6%

Table 2: Hitachi iQ Performance

Hitachi Vantara
A Hitachi Group Company

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
hitachivantara.com | community.hitachivantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
hitachivantara.com/contact

HV-BTD-DS-Hitachi-iQ-with-NVIDIA-HGX-16Oct24-A