Hitachi iQ

FAQ

Product Solutions Marketing

Revised November 2024

Hitachi Vantara Confidential - for approved partner use and external use under NDA

This FAQ document is designed to help you understand Hitachi Vantara's announcement and the latest developments within the Hitachi iQ brand portfolio.

What is the purpose of the AI / Hitachi iQ announcement?

November 2024 Announcement:

This announcement focuses on the general availability of Hitachi iQ with NVIDIA HGX™ (H100 and H200) GPU processors in an end-to-end solution stack of fast, scalable, and reliable resilient infrastructure for modern AI systems. Hitachi iQ with NVIDIA HGX provides cost-optimized solutions, tailored to a variety of industries and use cases, including large language models, inferencing, model training, analytics, digital twins, and image processing.

The latest offering from Hitachi iQ is designed to maximize AI performance and reliability, to deliver faster time to insight. Of note:

- Enhanced data processing of distributed file systems allows for more effective pattern identification while access to diverse datasets helps AI models learn from a wider range of scenarios.
- Hitachi iQ with NVIDIA HGX improves the economics and sustainability of AI through the use of zero-copy architecture that eliminates wasteful data copying and transfer times between storage silos for different phases of AI.
- High-density object storage allows scaling of compute and storage independently with the flexibility to adapt to diverse and fluctuating data sizes, data types, and workloads.
- Hitachi iQ with NVIDIA HGX combines storage, networking, servers and NVIDIA AI Enterprise™, an end-to-end, cloud-native software platform that accelerates data science pipelines and streamlines development and deployment of production-grade co-pilots and other generative AI applications.

July 2024 Announcement Recap:

The July 2024 announcement builds on our March 2024 press release and showcases the rapid progress Hitachi Vantara has made this year.

This July 2024 announcement highlights the tangible offerings that are now available to the market as a result of our steadfast execution of the strategy we outlined in March. This announcement features the following:

1. NVIDIA BasePOD certification of Hitachi Content Software for File with the NVIDIA DGX platform, allowing us to deliver the highest performing AI solution on the market.
2. New AI Discovery Service for Hitachi iQ to help customers accelerate their AI projects.

March 2024 Announcement Recap:

- Hitachi Vantara press release announcing partnership with NVIDIA and introducing Hitachi iQ, Hitachi Vantara's new portfolio of AI-ready infrastructure and solutions to accelerate and simplify customers adoption of AI. The press release was also promoted on Hitachi ltd.
- Hitachi press release announcing it's collaborating with NVIDIA to accelerate social innovation and digital transformation by combining Hitachi's domain solutions in operational technology (OT) and leadership in key industries including energy, mobility, and connected systems with NVIDIA's expertise in generative AI.

What is the summary of the AI / Hitachi iQ announcement and what are the key messages?

This announcement introduces Hitachi iQ with NVIDIA HGX^{TM} and includes 3 core announcement components:

1. **AI Ready Infrastructure**

- PROBLEM: Meet AI performance, scale, and reliability demands
- RESOLUTION: Hitachi iQ for NVIDIA HGX^{tm} accelerates performance through use of distributed file systems with the latest performance capabilities available to saturate GPUs with data and get the most value out of your processor spend.

2. **Reliability, Sustainability, and Efficiency**

- PROBLEM: AI demands more than just high performance. The cost and mission-criticality of AI systems means they must make efficient use of data, storage, power, cooling and floor space while maintaining maximum uptime.
- RESOLUTION: Hitachi iQ for NVIDIA HGX^{tm} improves economics and sustainability of AI through use of higher performance systems, zero-copy architecture, and high-density object storage. It allows scaling of compute and storage independently with the flexibility to adapt to diverse and fluctuating data sizes, data types, and workloads. And it maintains availability for continuous productivity with Hitachi's legendary reliability and support.

3. **Speed time to value for AI projects**

- PROBLEM: AI infrastructure is complex
- RESOLUTION: Hitachi iQ for NVIDIA HGX^{tm} comes pre-built by Hitachi Vantara for ease of deployment, operation, and maintenance through single vendor approach. Add NVIDIA AI Enterprise to speed deployment of AI models and workloads.

Why is this important?

- AI Infrastructure and Analytics is a huge market. Currently Hitachi does not currently address the larger AI infrastructure market.
- Speed time to insight by creating predefined solutions for markets of Hitachi expertise and interest. Hitachi iQ provides an “easy button” for AI solutions, creating an AI solution that is 60%-70% complete, injected with Hitachi engineering IP, enabling services to take the solution to 100% filling customer bespoke requirements.
- Hitachi cross business unit created AI solutions for industries that Hitachi and its global ecosystem of partners excel in, including manufacturing, energy, finance, transportation, and health sciences. These solutions will be branded and represented as Hitachi iQ for <insert solution name>.
- Hitachi cross Business Unit created AI solutions for industries with significant Hitachi footprint. Solutions will be branded as Hitachi iQ for <insert solution name>
- HV/Partner co-created AI solutions through our global solutions integrator (GSI) and solutions integrator (SI) partnerships, including AI discovery services delivered by Hitachi company partners Global Logic & Hitachi Digital. Solutions will be branded as Hitachi iQ for <*insert solution name*>.

What is Hitachi iQ?

Hitachi iQ is the single brand for all Hitachi AI related offerings which include AI-ready infrastructure, solutions and services. Hitachi iQ consists of purpose-built infrastructure, supporting the high-performance workloads necessary for successful AI, including storage, compute, & software.

Additionally, iQ is made up of AI solutions built by Hitachi & Hitachi Vantara alliance partners to address real-world AI use-cases.

The outcome from Hitachi iQ is a simplified way to provide AI solutions to our customers, while aligning with industry trends, and providing the latest in AI infrastructure.

What is the overall message for the Hitachi iQ

Hitachi iQ is Hitachi Vantara's new portfolio of infrastructure and solutions to address the AI market. Solving for infrastructure, Hitachi has combined the best in market offerings from NVIDIA with our award-winning high performance parallel filesystem, Hitachi Content Software for File (HCSF). Unlock the fastest time to insights with this powerful combination and accelerate GPU workloads by up to 20x, delivering unmatched efficiency and performance. Paired with Hitachi's object storage platform, HCP provides best-in-class economics for AI platforms, especially at scale.

In addition to Hitachi's infrastructure offerings, Hitachi will leverage our 110-year pedigree of powering Operational Technology Industries including Transportation, Energy, & Manufacturing to curate a portfolio of AI solutions offered within the Hitachi iQ portfolio. Built on NVIDIA's AI Enterprise framework, these solutions will be designed to simplify the adoption of AI driven outcomes across industry, furthering Hitachi's mission for improved sustainability & social innovation.

By working with both our global ecosystem of partners and our Hitachi sister companies, the goal for Hitachi iQ goal is simple. Hitachi Vantara wants to create market tailored AI solutions designed for both industries and businesses committed to AI emergence. The mission is to empower customer organizations to automate their processes, accelerate time-to-insights, and unlock innovation, allowing customers to reach their full potential.

What is the elevator pitch for Hitachi iQ?

What it is and who it's for: Hitachi iQ is an industry-optimized AI solution suite for organizations investing in AI/ML, GenAI, and other demanding GPU-driven workloads.

The problem it solves and its' value: Organizations are constantly looking for ways to automate their business, optimize time to market and develop new insights, products or innovations to propel their business forward. Hitachi iQ allows organizations to automate and accelerate their business through intelligent, performant, scalable and flexible GenAI, solutions and services in a hybrid cloud environment.

How it's Different: Unlike other approaches on the market today, Hitachi iQ goes beyond basic integration and testing by layering industry specific capabilities on top of the AI solution stack, so outcomes can be more specific and relevant to an organizations business.

What problem does it solve and what value does this bring to customers?

Most organizations are now racing to figure out how they can leverage, integrate, augment, or adopt AI. With the popularity and growth of tools like ChatGPT, Microsoft CoPilot, etc., organizations have quickly extrapolated how they can apply this kind of technology to do things like automate operations, improve customer service, increase productivity, speed product delivery, reduce defects, the list goes on. The sad reality is that many AI projects fail. Sometimes it's ill-defined goals of the project itself or lack of expertise or skills with AI, and other times it's the technology side of the equation where the technology is not able to deliver as anticipated. Hitachi's approach is to help customers be successful in their AI journey by delivering an AI-ready infrastructure stack, which can handle the toughest of AI workloads and AI-ready solutions and services that take advantage of Hitachi's 110+ years of combined IT & OT experience to create industry and use case specific solutions targeted to deliver specific customer outcomes.

With Hitachi iQ, customers can:

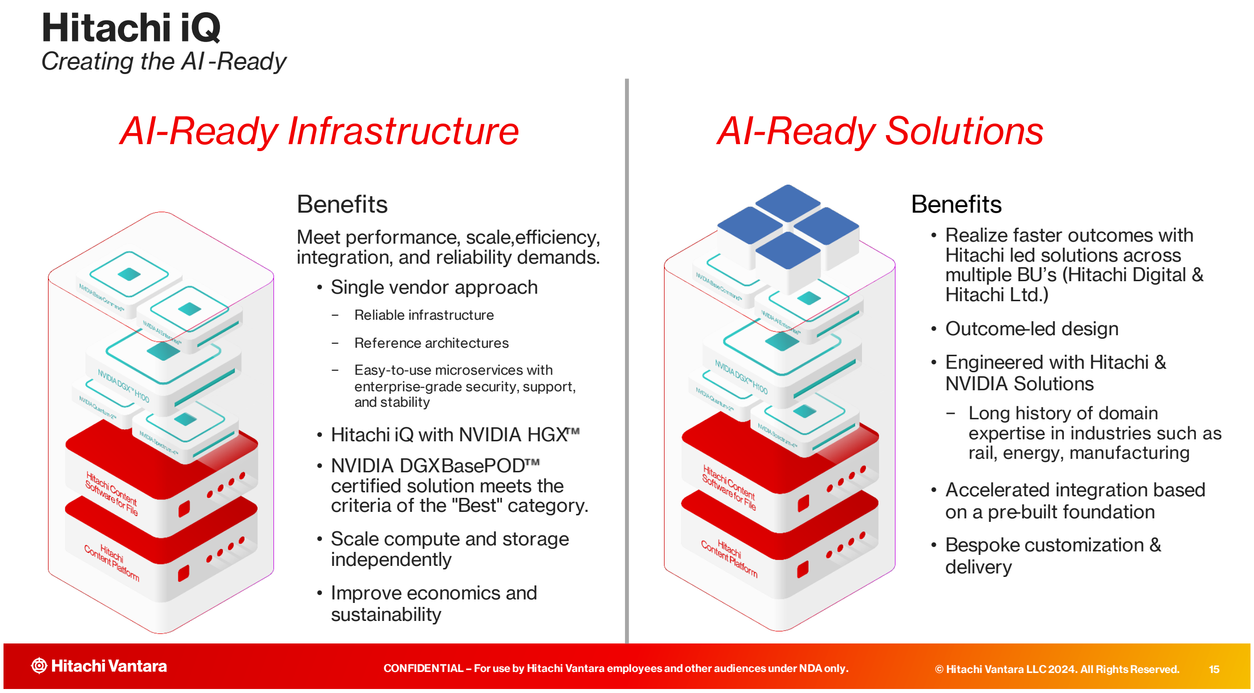
- **PERSONALIZE THEIR AI:** Built to meet the rigors of AI, engineered for industry outcomes, but customized to their needs "Your AI Your Way".
- **ACHIEVE FASTER INSIGHTS:** Accelerated architecture delivers massive performance starting at 600GB/s and 22M IOPs resulting in as much as 20X reduction in GPU processing time through better utilization of resources.
- **SIMPLIFY TO SCALE:** Validated designs and solution blueprints provide the flexibility and scale to rapidly develop, test, and deploy modern applications while adapting to fluctuating customer workload demands.
- **LOWER TCO:** Leverage lower cost, erasure coded, scale out object storage to safeguard data, catalog for re-use and store data long term while optimizing costs
- **IMPROVE ACCURACY:** Increase quantity and quality of data to improve reliability of results. Identify, classify, transform, move, consolidate and prepare data to get the most value out of AI/ML initiatives.

What is Hitachi iQ?

Hitachi iQ is the single brand for all Hitachi AI related offerings. iQ consists of purpose-built infrastructure, supporting the high-performance workloads necessary for successful AI, including storage, compute, & software.

Additionally, iQ is made up of AI solutions built by Hitachi and Hitachi Vantara partners to address real-world AI use-cases.

The outcome from Hitachi iQ is a simplified way to provide AI solutions to our customers, while aligning with industry trends, and providing the latest in AI infrastructure.



Hitachi iQ

Creating the AI -Ready

AI-Ready Infrastructure

Benefits

Meet performance, scale,efficiency, integration, and reliability demands.

- Single vendor approach
  - Reliable infrastructure
  - Reference architectures
  - Easy-to-use microservices with enterprise-grade security, support, and stability
- Hitachi iQ with NVIDIA HGX™
- NVIDIA DGXBasePOD™ certified solution meets the criteria of the "Best" category.
- Scale compute and storage independently
- Improve economics and sustainability

AI-Ready Solutions

Benefits

- Realize faster outcomes with Hitachi led solutions across multiple BU's (Hitachi Digital & Hitachi Ltd.)
- Outcome-led design
- Engineered with Hitachi & NVIDIA Solutions
  - Long history of domain expertise in industries such as rail, energy, manufacturing
- Accelerated integration based on a pre-built foundation
- Bespoke customization & delivery

Hitachi Vantara   CONFIDENTIAL – For use by Hitachi Vantara employees and other audiences under NDA only.   © Hitachi Vantara LLC 2024. All Rights Reserved.   15

What are the key components and capabilities of Hitachi iQ with NVIDIA HGX?

- **AI Software Stack**
  - Enterprise-class platform for AI training
  - End-to-End software for accelerated AI pipelines
  - Production grade pretrained AI models
  - Cluster provisioning, workload management, and infrastructure monitoring (DGX): NVIDIA Base Command Manager
  - AI Framework / Management: NVIDIA AI Enterprise - standardize the toolkit and prepare for evolving needs with NVIDIA AI Enterprise
- **Hardware**

Items	Specs
CPU	Dual Socket 5th Gen Intel Xeon 48C or 56C CPU
Memory	32 x 64GB DDR5-5600 2Rx4 ECC RDIMM
GPU Options	8 x NVIDIA HGX H100, SXM5 8 x NVIDIA HGX H200, SXM5
Networking	8 x NVIDIA PCIe 1-port IB 400GE OSFP or 8 x NVIDIA PCIe 1-port IB NDR200 OSFP for HCSF 2 x 10GB RJ45 for Managent server
Storage	OS: 2 x 960GB NVMePCIeGen4 M.2 Optional Storage: 16 x 2.5” NVMe Drives up to ~245TB
Power Supply	6 x AC-DC 3000W, Titanium efficiency as Default. Optional 2 x AC-DC 3000W, Titanium efficiency for upgrade GPU
Optional Networking	3 x 10/100 GbE and 200/400 GbE/IB NIC
Management	Redfish API/IPMI2.0

- **High-Performance Parallel Filesystem**
  - Primary: Hitachi Content Software for File with GEN5 storage
  - Accelerated filesystem to match the mixed AI workload
  - Modern consumption improving AI workloads (GPUDirect, POSIX, S3)
  - Latest generation platform to support the highest concentrated workloads
  - Native data offloading capabilities to support better economics at scale
- **Scale-out Object Storage Tier**
  - Optional object / tiering / data protection: Hitachi Content Platform (HCP) & Hitachi Content Platform for Cloud Scale (HCP:CS)
  - Object storage density to provide best storage economics at density
  - Erasure coding protection designed for large scale volumes of data
  - Data reduction & compression for ultimate capacity efficiency

What are the key components and capabilities of Hitachi iQ with NVIDIA DGX?

- **AI Software Stack**
  - Enterprise-class platform for AI training
  - End-to-End software for accelerated AI pipelines
  - Production grade pretrained AI models
  - Cluster provisioning, workload management, and infrastructure monitoring (DGX): NVIDIA Base Command Manager
  - AI Framework / Management: NVIDIA AI Enterprise - standardize the toolkit and prepare for evolving needs with NVIDIA AI Enterprise
- **Accelerated Compute**
  - Fully integrated hardware and software solution on which to build your AI Center of Excellence
    - NVIDIA DGX H100
    - Powered by eight NVIDIA H100 Tensor Core GPUs per node
    - This is offered through NVIDIA BasePOD Certification & is specifically a meet in the channel compute offering
    - Hitachi HGX H100 and H200
    - PCI-e/Midrange L40S, H100, H200
  - Proven reference architectures for AI infrastructure delivered with Hitachi's leading storage technology
  - Flexibility to support both customers that want and partners who sell DGX or as an alternative to the DGX offering, customers and partners can choose the Hitachi HGX offering as well.
- **Switching**
  - Highly scalable & self-healing network capabilities
    - NVIDIA QM9700 400Gb/s InfiniBand Switches
    - NVIDIA SN4700 400GbE Ethernet Switches
    - All necessary cables & optics to support connectivity
  - High-speed, extremely low-latency, and scalable networking

- **High-Performance Parallel Filesystem**
  - Primary: Hitachi Content Software for File with GEN5 storage
  - Accelerated filesystem to match the mixed AI workload
  - Modern consumption improving AI workloads (GPUDirect, POSIX, S3)
  - Latest generation platform to support the highest concentrated workloads
  - Native data offloading capabilities to support better economics at scale
- **Scale-out Object Storage Tier**
  - Optional object / tiering / data protection: Hitachi Content Platform (HCP) & Hitachi Content Platform for Cloud Scale (HCP:CS)
  - Object storage density to provide best storage economics at density
  - Erasure coding protection designed for large scale volumes of data
  - Data reduction & compression for ultimate capacity efficiency

What Is BasePOD?

BasePOD is an integrated solution consisting of NVIDIA hardware and software components, MLOps solutions, and third-party storage. Leveraging best practices of scale-out system design with NVIDIA products and validated partner solutions, customers can implement an efficient and manageable platform for AI development. The designs in this BasePOD reference architecture (RA) support developer needs, simplify IT manageability, and infrastructure scaling from two nodes to dozens with certified storage platforms from an industry-leading ecosystem. Optional MLOps solutions can be integrated with BasePOD to enable a full stack solution to shorten AI model development cycles and speed the ROI of AI initiatives.

The BasePOD certification tests the capabilities of the storage system, Hitachi Content Software for File, to ensure the performance & reliability requirements for AI systems is met. Additionally, if offers a seamless support pathway for customers integrating the Hitachi Content Software for File storage system with their GPU platform, ensuring comprehensive end-to-end support for the entire solution.

What can I sell and when?

AI-Ready Infrastructure

Today, Hitachi and Hitachi partners have the ability to drive the Hitachi iQ with Nvidia HGX solution, inclusive of Nvidia AI Enterprise, or sold separately for general purpose AI use cases.**Hitachi iQ with NVIDIA HGX:** For organizations and researchers who require flexibility, scalability, and a platform tailored to their specific needs, the Hitachi iQ with HGX is a great choice. Hitachi partners can sell a complete HGX solution to customers.

1) certified by NVIDIA to sell DGX for the compute portion of a Hitachi iQ solution.
2) **NVIDIA AI Enterprise** for GPU workloads can be added to HGX orders or can be sold alone HCSF configurations.
   - By reselling NVIDIA AI Enterprise, Hitachi and Hitachi partners empower businesses of all sizes with access to the industry's leading AI software suite, designed to accelerate deployment, management, and scaling of AI solutions.
   - Comprehensive Tools: NVIDIA AI Enterprise offers a complete suite of AI and data analytics tools that streamline the AI development lifecycle.
   - Ease of Integration: Clients benefit from seamless integration with existing IT environments, reducing complexity and accelerating time-to-value.
   - Proven Performance: Backed by NVIDIA's extensive expertise, these solutions are optimized for performance and reliability across various industries.

AI-Ready Solutions

Hitachi Vantara will also be bringing industry and use cases specific solutions to market by packaging additional software, tools, services, etc. with Hitachi AI ready stacks to deliver more customized and targeted outcomes for our customers. For example,, these solutions will be named as follows for pre-built AI solutions: Hitachi iQ for *<market solution>*

1. Hitachi iQ for Fraud Detection
2. Hitachi iQ for Manufacturing Insights
3. Hitachi iQ for Fleet Management
4. Hitachi iQ for Energy

Hitachi AI Center of Excellence: Provided by Hitachi Vantara's Center of Excellence (COE), this is an Incubation Space for Hitachi employees, partners, and customers to research and test AI solutions while expanding Hitachi's expertise in the AI market. Additionally, the COE is used as a space for continuous improvement of technology offerings including performance characterization, vetting new technologies, and evolving Hitachi and partner capabilities.

How do I order Hitachi iQ?

For HGX opportunities, Hitachi can sell the whole solution stack, starting with Hitachi Content Software for File for primary storage and including HGX H100 nodes, management servers, networking fabric components, and management software.   To build a quote for HGX, use “Hitachi iQ Product Sales” in SFDC/CPQ. Building a Hitachi iQ solution begins with choosing the ‘Hitachi iQ w/HGX’ solution heading, front-end networking, storage networking, and power (all of which currently have a single default entry). From there, HGX H100 nodes, management server(s), Hitachi Content Software for File, and other constituents can be added and configured with desired choices for processors, memory, drives, GPUs, switches, software, and other components. Professional services and maintenance options can also be configured and added to the complete solution as part of the process. Note that it is recommended to check “Make your selections and check this to build the configuration” on the main page when you are ready to output a finished configuration (and to un-check and re-check when making changes to force re-calculations).

How do I get help with configuring a full Hitachi iQ with HGX solution?

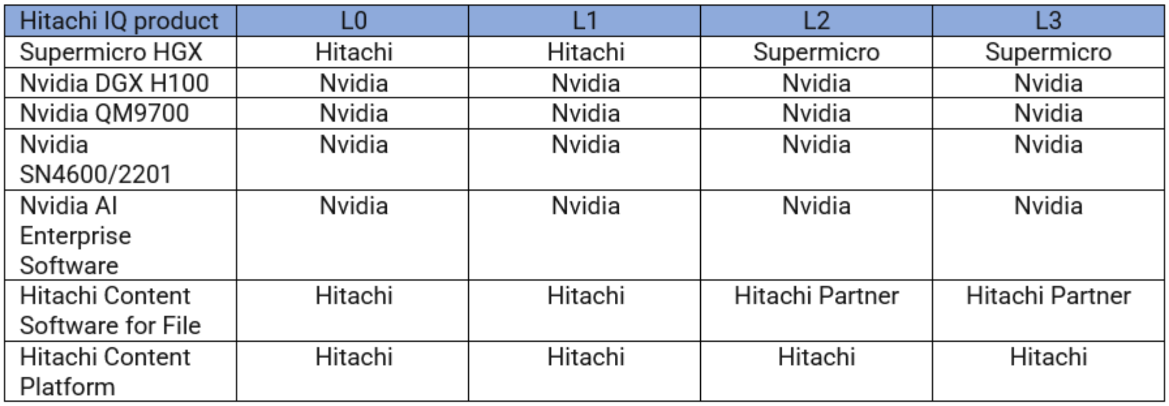
The Hitachi iQ HGX configurations can be complex so Hitachi is here to help. Please reach out to AI.Incubation.Team@hitachivantara.com for assistance.

What is the nature of the Hitachi Vantara Partnership with NVIDIA?

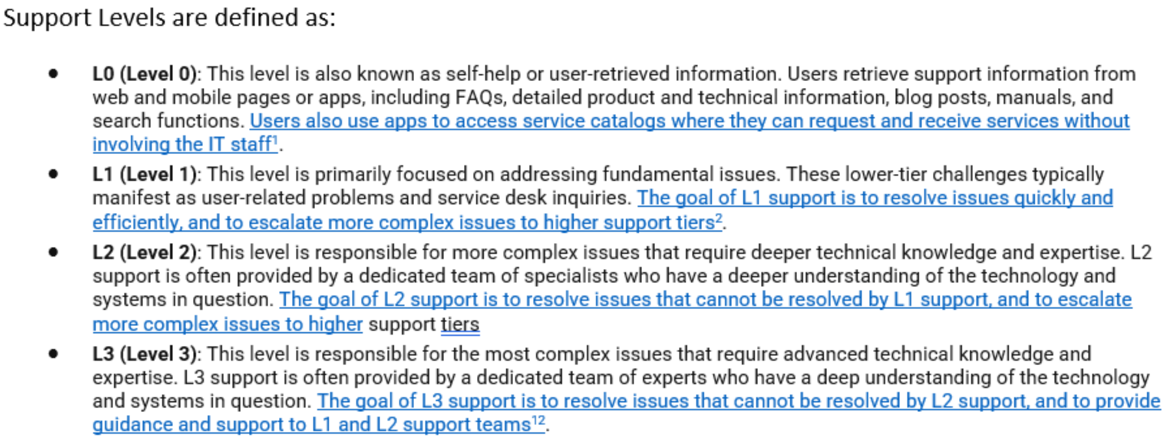
- Hitachi Vantara is an NVIDIA preferred partner and a part of their OEM, Solution Advisor and Solution Integration partner programs. Hitachi Vantara also holds competencies of Compute, Network, DGX, HGX and AI.

How is the Hitachi iQ solution supported by Hitachi Vantara?

As mentioned earlier in the FAQ, the Hitachi iQ solution consists of several products. The Hitachi support and services team provides various support options based on the specific product. The table below defines the support levels and services for each of the component products:



Hitachi IQ product	L0	L1	L2	L3
Supermicro HGX	Hitachi	Hitachi	Supermicro	Supermicro
Nvidia DGX H100	Nvidia	Nvidia	Nvidia	Nvidia
Nvidia QM9700	Nvidia	Nvidia	Nvidia	Nvidia
Nvidia SN4600/2201	Nvidia	Nvidia	Nvidia	Nvidia
Nvidia AI Enterprise Software	Nvidia	Nvidia	Nvidia	Nvidia
Hitachi Content Software for File	Hitachi	Hitachi	Hitachi Partner	Hitachi Partner
Hitachi Content Platform	Hitachi	Hitachi	Hitachi	Hitachi



Support Levels are defined as:

- **L0 (Level 0)**: This level is also known as self-help or user-retrieved information. Users retrieve support information from web and mobile pages or apps, including FAQs, detailed product and technical information, blog posts, manuals, and search functions. Users also use apps to access service catalogs where they can request and receive services without involving the IT staff[1].
- **L1 (Level 1)**: This level is primarily focused on addressing fundamental issues. These lower-tier challenges typically manifest as user-related problems and service desk inquiries. The goal of L1 support is to resolve issues quickly and efficiently, and to escalate more complex issues to higher support tiers[2].
- **L2 (Level 2)**: This level is responsible for more complex issues that require deeper technical knowledge and expertise. L2 support is often provided by a dedicated team of specialists who have a deeper understanding of the technology and systems in question. The goal of L2 support is to resolve issues that cannot be resolved by L1 support, and to escalate more complex issues to higher support tiers
- **L3 (Level 3)**: This level is responsible for the most complex issues that require advanced technical knowledge and expertise. L3 support is often provided by a dedicated team of experts who have a deep understanding of the technology and systems in question. The goal of L3 support is to resolve issues that cannot be resolved by L2 support, and to provide guidance and support to L1 and L2 support teams[12].

What makes Hitachi's approach to AI unique?

Only Hitachi Ltd. has the in-depth industrial expertise across energy, transportation, and more to utilize GenAI to accelerate digital transformation, customer experience, and people’s lives!

Through Hitachi Ltd.'s extensive legacy in improving society, these companies are fueling true industrial impact. Now armed with Hitachi iQ, Hitachi & NVIDIA's impact on society is being extended to fuel further social innovation.

With Hitachi Ltd and Hitachi Vantara, NVIDIA is now able to reach and accelerate innovation across numerous industry verticals.

Unlike other storage vendor engagements with NVIDIA which have focused on general-purpose AI and analytics appliances, Hitachi Vantara is able to harness the vast industrial market knowledge and solution penetration of Hitachi Ltd. companies.

What customers should be targeted for these solutions?

TARGET PERSONAS

	EXECUTIVES	DATA SCIENTISTS AND RESEARCHERS	SOFTWARE ENGINEERS / DEVELOPERS	IT INFRASTRUCTURE AND OPERATIONS MANAGERS
Cares About	Return on investment (ROI), aligning AI investment to business strategy	Optimizing data analysis, reducing time to insights	Building and optimizing applications, automating work	Building reliable IT infrastructure with minimal footprint
Title	CEO, CTO, President, VP of Data Science, VP of Engineering, VP of Analytics	Data Analyst, Data Engineer, Data Architect, Research Scientist, Research Associate, Professor	Software Developer, Software Engineer, Programmer, DevOps Engineer	IT Infrastructure Manager, IT Operations Manager, Infrastructure Engineers

See Sales Quick Reference Guide on Partner Portal for more detailed information.

Which use cases should I target with Hitachi iQ? (e.g., Deep Learning training, Large Language Model, digital twin, Genomic mapping, seismic research, etc.)

The possibility of AI and Generative AI in the market is endless. The industry is using this for everything from creating art and graphic design elements to writing code and generating marketing tag lines. The usecases are endless.

These popular usecases can be used to provide immediate benefit to any organization.

Whether customers are just getting started or are far along in their AI journey, these generalpurpose use cases automate business processes and operations and deliver tailored customer experience.

Example Solutions

Customer Service Voice Assistant: Customer services assistance revolutionizes customer service with its advanced voice recognition technology, offering real-time responses and personalized assistance. These virtual assistants streamline interactions to ensure customer queries are handled efficiently and with a human touch.

LLM Recommender System: Leveraging the power of Large Language Models, recommender systems deliver highly accurate, context-aware suggestions to enhance user experiences across digital platforms. It personalizes content, products, and services, making discovery seamless and engaging.

Coding & Development Copilot: Copilots can act as an invaluable partner in the development process, offering real-time suggestions, debugging assistance, and code optimization. It accelerates development cycles and enhances code quality by leveraging advanced AI algorithms.

Automated Document Processing and Analysis: This use case applies AI to streamline the processing, analysis, and management of large volumes of documents within enterprise environments. By automating tasks such as data extraction, categorization, and summarization, businesses can significantly reduce manual workloads, improve accuracy, and enhance decision-making processes across various departments like finance, HR, and legal.

Financial Reporting & Accounting: By utilizing large language models that specialize in tax, accounting, cash flow, and more, these solutions can reduce the repetitive tasks that works and consultants are required to do, helping streamline operations and reduce errors for data entry, transaction categorization, and invoice processing.

Edge Inference: Edge Inference technology 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.

Who should I engage if my customers express interest after the announcement, and how?

Hitachi Vantara AI Incubation Sales and Technical Specialists

AI.Incubation.Team@hitachivantara.com

Partners & Alliances

AI Global Alliance Manager - NVIDIA, Weka, and Supermicro: Christine Brennan

Christine.brennan@hitachivantara.com

Channel Partner Readiness: Rob Williams

PartnerPrograms@HitachiVantara.com or Rob.Williams@HitachiVantara.com

What assets and collaterals will be available now and in the future?

- **External**
  - Hitachi iQ Solution Profile
  - AI Discovery Service for Hitachi iQ Solution Profile
  - Hitachi iQ elevator pitch and teaser video
  - 2 Hitachi videos from NVIDIA GTC
    - The Hitachi AI Difference
    - Evolving AI with Hitachi
  - Blog by Octavian and related blog series to follow
  - Hitachi iQ icons (available for download from the brand portal)
  - Hitachi iQ NVIDIA DGX BasePOD reference architecture
  - Hitachi iQ HGX reference architecture white paper
  - ESG eBook summarizing the results of the Hitachi Vantara sponsored market survey on Enterprise Infrastructure for Generative AI and the press release promoting the results of the study
  - Top 10 Reasons: Hitachi iQ
  - AI Analytics webpage
  - Hitachi iQ webpage
  - Hitachi channel partners can use the Partner Portal and will receive updates and information through the partner newsletter.
  - AI for Dummies book (*future*)
  - *And more to come...*

What resources exist for partner awareness, enablement, and marketing? Please visit https://partnerportal.hitachivantara.com/s/hitachi-iq for sales and technical assets, and https://partnermarketing.hitachivantara.com/#/page/welcome for customizable campaigns partners can use to complement your current lead generation initiatives.

Is there a certification program available to us as an NVIDIA partner?

Self-paced NVIDIA training is available free of charge. Contact Christine.brennan@hitachivantara.com to get access to NVIDIA training modules.

What solutions will Hitachi iQ be competing against?

The vendors with solutions in this space are, DDN, DELL, Hammerspace, IBM, NetApp, Pure Storage, and VAST Data.

DDN A3i

- DDN's appliance solution is based on Lustre, a 25-year-old parallel filesystem known for its management complexities, its lack of stability at massive scale, and performance inefficiencies when confronted with mixed data profiles.

DELL PowerScale

- DELL PowerScale is the newest name of the solution EMC acquired when it added Isilon Systems to its portfolio of storage asset. PowerScale, while known for its performance and scalability, does have several limitations.
  - Scale and Performance may be compromised if customers exceed 80% of the total capacity. When it hits 85% the performance degradation becomes more pronounced, at 90% or higher, significant performance degradation is noticeable, and operational disruptions are almost certain. DELL recommends planning to buy additional capacity when the cluster reaches 75%.
    - Hitachi iQ outperformed DELL Powerscale in NVIDIA's BasePod certification testing 4x read and 2x write performance with nearly 50% of the infrastructure.
  - In AI environments requiring real-time data processing and high throughput, PowerScale may encounter performance bottlenecks, especially with data profile consisting of small and large files.

Cost can be an issue due to its high initial investment, and due to the management overhead, which can be complex and resource intensive. Some customers may find they are in an "under-sized" cluster due to inaccuracies in gauging the data efficiency ratio for their data, resulting in the need to expand.

Hammerspace

Hammerspace has entered the market under the cloud of "partnership" and not as a competitor. They offer global data management and orchestration, providing a valuable platform for organizations working with large-scale unstructured data. However, while their global data management may prove useful to some working with large-scale unstructured data, Hammerspace does not perform well when applied to AI and GenAI workflows. These limitations can impact performance, scalability, and ease of integration with AI frameworks.

Data Latency and Performance

- While Hammerspace provides global data access, performance in distributed environments can suffer due to network latency, which is critical for AI workloads that require rapid data processing for real-time applications.

Scalability and Infrastructure

- Although it supports horizontal scaling, Hammerspace may struggle with the extreme computational demands of large GenAI models, especially in terms of storage throughput and compute-resource alignment.

Complexity in Data Management

- Managing vast and dynamic datasets for AI training can become cumbersome due to the complexity of orchestration and the need for seamless integration with AI frameworks.

Integration with AI Frameworks

- Direct integration with popular AI frameworks (e.g., TensorFlow, PyTorch, and Hugging Face) may require custom configurations, potentially adding complexity and delays in AI development cycles.

Data Gravity and Transfer Costs

- Transferring large datasets globally can be expensive and time-consuming, especially when balancing cloud, on-premises, and edge environments.

IBM

IBM Spectrum Scale (formerly GPFS) is over 25 year old parallel file system originally designed for video on-demand streaming. It has evolved to support HPC environments but has some limitations when considering it for AI and GenAI workloads.

Complexity

- Spectrum Scale is very complex and difficult to set up and managed day-to-day. Since it was originally designed for large file data characteristics, organizations may experience bottlenecks and additional management overhead when AI workflows require frequent changes to support small, large or a mix of these file types.
- The complexities of management and configuration increases the operational burden, particularly for teams focusing on AI development rather than infrastructure management.

Latency and Throughput Bottlenecks

- Slower data access can lead to longer training times, which is critical when training large AI models. Spectrum Scale does suffer from having multiple MDS (Meta Data Servers) when not configured properly. This can lead to data access and performance issues and latency due to metadata synchronization between MDS nodes.
- Metadata hotspots are a real issue with Spectrum Scale, where one or a few MDS nodes handle a disproportionate amount of the metadata requests. This causes contention and performance degradation

NetApp

- NetApp ONTAP has been in the market since the early 1990s and was the early standard for enterprise IT file share services. However, its architecture lacks the high-performance characteristics many of the more demanding AI and Generative AI workloads require, such as large-scale training, and deep learning models. As customers are entering the realm of GenAI, the performance of NetApp may be "good enough" but as the customer's performance needs scale, they will quickly outgrow what NetApp is able to offer within the ONTAP family. Alternatively, NetApp does offer BeeGFS on its all-flash E-Series disk. BeeGFS has a long list of limitations as well, such as:
    - No support of data protection such as erasure coding, no native file encryption
    - Installation and Management Complexities
    - The metadata servers may become a bottleneck and impede performance as the number of files and operations increases
    - BeeGFS claims low latency, high bandwidth, however, due to the significant network overhead BeeGFS generates to achieve customer expectations it may require the purchase of extreme high-bandwidth, low-latency network infrastructure to overcome its overhead.
    -

PURE Storage AiRi

- Pure Storage AiRi is based on its FlashBlade //S product offering, and like NetApp, lacks the performance characteristics for high-end AI and GenAI application demands such as large-scale training and deep learning models. While Pure Storage has claimed FlashBlade //S is 3-4x faster than the previous FlashBlade generation, it does not publicly share any benchmarks for customers to compare.

VAST Data

VAST Data is a scalable storage platform designed to meet the needs of modern data applications, including AI and GenAI workloads. It leverages an architecture focused on disaggregated storage and computational resources, which can provide significant advantages in AI workflows. However, it has some significant limitations when applied to AI and GenAI environments

Cost Efficiency at Scale

VAST Data's architecture, which combines flash storage with caching to NVMe (Storage Class Memory), can become expensive when deployed at a very large scale. The cost of using this type infrastructure may be prohibitive for organizations that need to store and manage petabytes or exabytes of data.

Hitachi iQ with Content Software for File offers simplicity, scale, and performance for the customer's most demanding GenAI workloads. With market leading read/write performance, extremely high IOPs, combined with the lowest latency, Hitachi iQ with Content Software for File will scale with the customer's diverse mix of data profiles, and application requirements. None of our competitors can offer the level of performance, worldwide support and integration offered by Hitachi Vantara.

If you are looking for sales enablement and training, please reach out to PartnerPrograms@HitachiVantara.com.

If you have any more questions, please reach out to your Hitachi Vantara distributor, partner account manager, or representative.