Cloudian HyperStore Data Platform Reference Architecture for NVIDIA GPU Platforms from Supermicro

Table of Contents

Executive Summary

As artificial Intelligence (AI) and Machine Learning (ML) workloads move into mainstream IT, a new way of managing and accessing data is required to support the goal of maximising the “value out of” data to enable accurate predictions and decisions. The AI & ML landscape can be complex with multiple data pipelines depending on data type and source, multiple data consumers demanding different data sets and/or formats and all the data pre-processing stages of transforming data from raw to consumable format.

Ingest – Transform – Analyse – Curate - Log

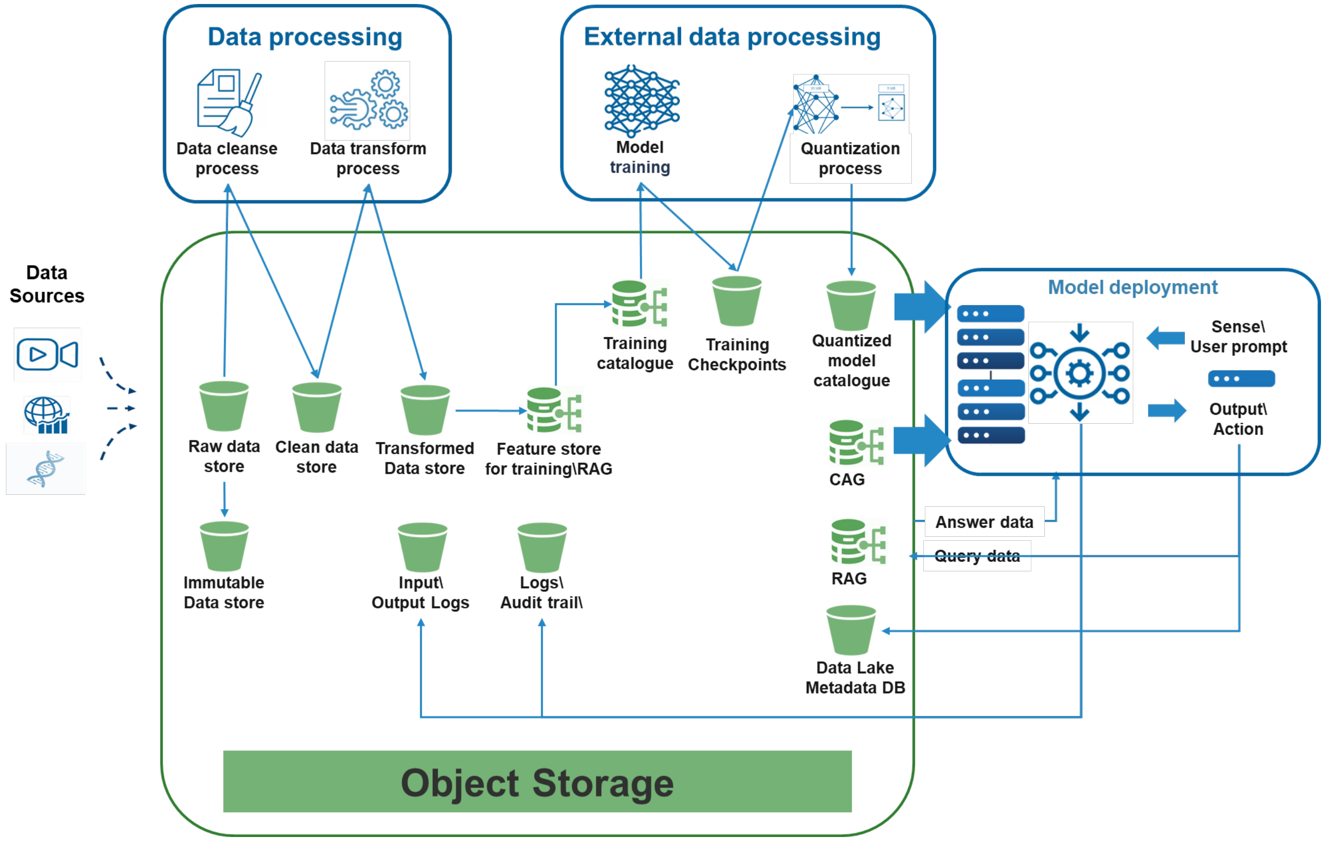
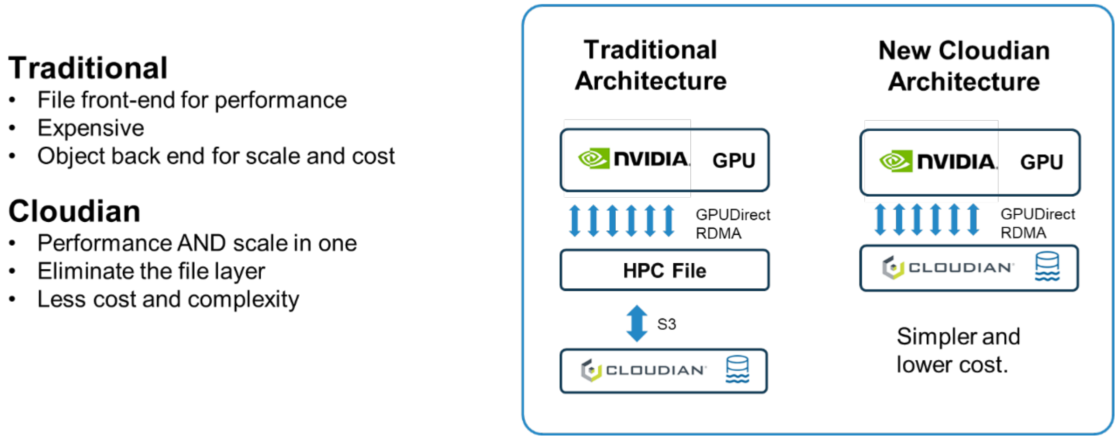


Figure 1: Complexity of AI & ML environments

At each stage a different IO workload profile is presented to the storage system, whether it’s storing newly generated raw data, cleaning data, transforming data from one format or storage method to another, ingesting data into ML training processes, storing training checkpoints or supporting AI models with external data sources for inferencing workloads. High performance capability for transferring large data files is crucial in these environments, especially in LLM training and inferencing applications.

Until now, the performance requirements of accelerated workloads such as AI and ML have typically been addressed with niche focussed scale out HPC storage solutions using client-based software to enable parallel NFS IO capabilities. This paper will show how the Cloudian HyperStore object storage platform can compete directly with HPC file solutions for even the most strenuous AI training and inferencing use cases.



Traditional
- File front-end for performance
- Expensive
- Object back end for scale and cost

Cloudian
- Performance AND scale in one
- Eliminate the file layer
- Less cost and complexity

Figure 2: Object storage platform reduces cost and complexity

This reference architecture will demonstrate, through performance testing, how a Cloudian HyperStore data storage cluster can scale both performance and capacity within a single managed system to deliver the performance required by these accelerated workloads at lower costs compared to having dedicated file servers for high-speed scratch storage access.

Beyond that, Cloudian HyperStore is more than just a bunch of fast flash storage drives that provide high performance access. Demands on the data platform serving accelerated compute workloads need to do so much more than just provide rapid access to data.

- **Centralized data storage and management**: eliminating data silos and making data easier to share
- **Scalability**: limitless and dynamic to keep up with the volume and velocity of data generated for AI use
- **Security:** Security, control, governance, and ownership of data is crucial and, in many cases, mandated.
- **Velocity** – Handling large datasets at high velocity for data provisioning and versioning
- **Reliability** – Data protection, built-in ransomware protection, and dynamic provisioning of storage
- **Cloud Adjacency** – Connect with resources across multiple DCs and clouds to load knowledgebase repositories
- **Efficiency** – Enterprise storage features such as compression, data lifecycle management on data sets
- **Secure Multi-Tenancy** – Adaptive QoS to isolate noisy neighbours
- **Data sovereignty:** On-prem data lake for regulatory compliance and control.
- **Diverse Data Management**: Capable of handling object and file data types, essential for AI versatility.

Machine Learning projects need to be **versionable**, **deployable** and **responsible and** can be automatically **retrained** and **monitored.** With extensive data curation capabilities facilitated by a scalable metadata database, tracking of relevant data is always on for the system to audit, rollback and replay as required at any stage of the pipeline.

Introduction

As computing capability increases through the use of new types of processing units (think GPU and DPU alongside CPU) and new networking technologies (InfiniBand, RoCE), the demand for faster access to more and more data has necessitated a rethink of how to keep these accelerated computing platforms fed with data to ensure maximum productivity and output accuracy within defined timescales. Data analytics datasets continue to increase in size and the time spent loading data directly impacts application performance.

NVIDIA GPUDirect Storage (GDS) is part of NVIDIA's Magnum IO software stack which enhances IO activity across the entire accelerated compute stack to simplify and speed up data movement, access, and management for multi-GPU, multi-node processing systems. Magnum IO supports NVIDIA CUDA-X™ libraries and makes the best use of a range of NVIDIA GPU and NVIDIA networking hardware topologies to achieve optimal throughput and low latency. CUDA is a parallel computing platform and programming model that allows developers to use GPUs for general-purpose processing.

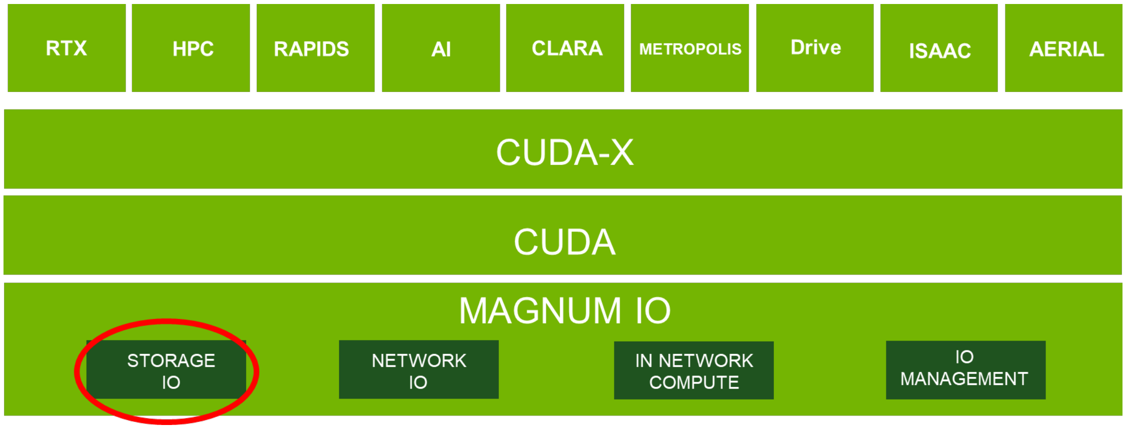


Figure 3: NVIDIA software stack

GDS enables an engine, in the host and storage platforms Network Interface Cards (NICs), to move data on a direct data path for direct memory access (DMA/RDMA) transfers between GPU memory and storage, without burdening the CPU avoiding a bounce buffer through the CPU. This direct path increases system bandwidth and decreases the latency and utilization load on the CPU. With the Parallel data transfer capability from multiple HyperStore nodes, RDMA data transfer accelerates data rates to ensure full GPU utilization. GDS APIs transfer data using one sided RDMA verbs (read/write) to allow the storage server to transfer the data directly to and from the remote GPU server's memory. This can be GPU or system memory.

RDMA can be used to load data directly into both GPU or CPU memory depending on the application architecture and use case. One typical performance trait of object storage has been the latency incurred using http protocol for data transfer. GDS removes these limits by separating the S3 control path (HTTP over TCP) from the data path, keeping the S3 authorization/authentication with standard protocol while accelerating the data path.

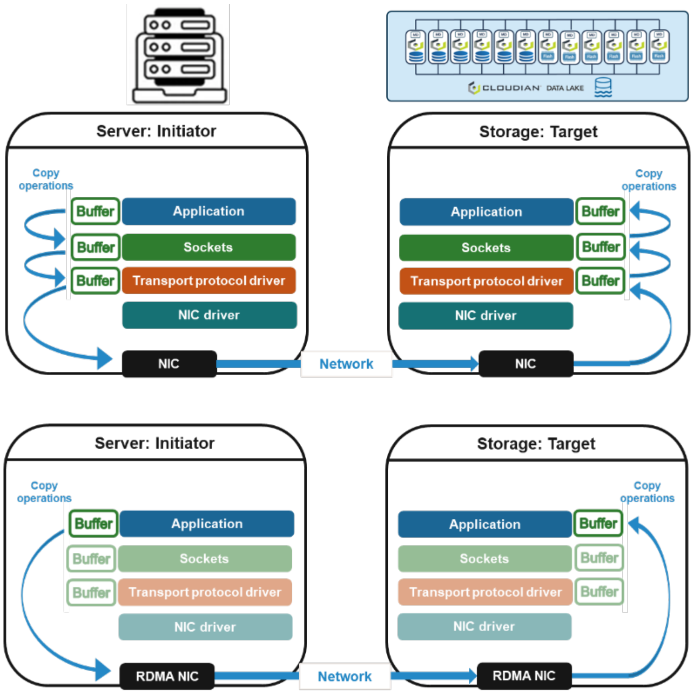


Figure 4: NVIDIA and Cloudian GPUDirect IO flow

Typically, an S3 application will make data requests such as S3 GetObject, S3 GetObjectRange, PutObject, which is processed by the HyperStore server and data sent back via http to the requesting client. Enabling GDS bypasses the typical operating system stack for data transfer. Requests are initiated via S3 API to move data between storage and GPU and instead of routing data through system memory via the CPU, GDS bypasses this step by RDMA transfer directly to GPU memory, minimizing latency.

To help overcome barriers for generative AI deployments, NVIDIA has introduced a partnership program for GPUDirect Storage (GDS) technology partners. Together Cloudian, Supermicro and NVIDIA combine to provide a single data storage platform that delivers both scalable performance and capacity.

This reference architecture document describes how to size Cloudian Hyperstore storage clusters to support AI model training and inference workloads built on HGX servers from Supermicro and NVIDIA GPUs, demonstrating how the solution can be scaled across larger compute clusters to support increasing complexity and scale requirements.

Cloudian HyperStore is a clustered scale out platform, which provides a single data storage namespace across multiple data storage locations providing site redundancy and options for a variety of edge to core deployments. As the infrastructure for AI workloads grow so do the data storage performance and capacity requirements. More GPUs will provide benefits such as speeding up training, inference and other GPU accelerated data analytics. Being able to feed GPUs data is the major consideration in any accelerated computing workflow as GPU idle time is expensive, keeping those GPUs fed with data and running is the aim of the game.

The architecture presented in this document can scale from relatively minimal deployments supporting 32 HGX or OVX servers to deployments hosting thousands of GPU servers.

Beyond the performance requirements, the capacity scale of the data storage platform is of equal importance. The reliance on good quality data is so important for AI training. As AI use cases evolve, organisations will have multiple varying data pipeline streams feeding different AI models for different use cases. There will be times

when some of the raw datasets are used within multiple model trainings and/or used as RAG data sources for providing domain knowledge expertise support to foundational models. Equally for training checkpoints and inference KV caches, these datasets can be reused in other environments\circumstances reducing cost of processing and improving overall inference response times by reusing pre-processed data

To support such a complex system, data curation becomes vital. Knowing where and what your data is and how to efficiently store and manage this data across all the data storage devices, clouds and data centres that organisations need to manage.

Data storage for AI and ML workloads

The complete end to end AI stack is complex, even at its most basic, so when you start to consider a multi-modal AI implementation with RAG and CAG data augmentation with multiple data feeds both batched and real time and inferencing feedback in real time to enrich raw data as it is first ingested… and then the world gets very complicated very fast!!

The end-to-end stack can be broken down into discrete activities that each have a different impact on a data storage platform depending on its processing requirements and the type of data objects being processed.

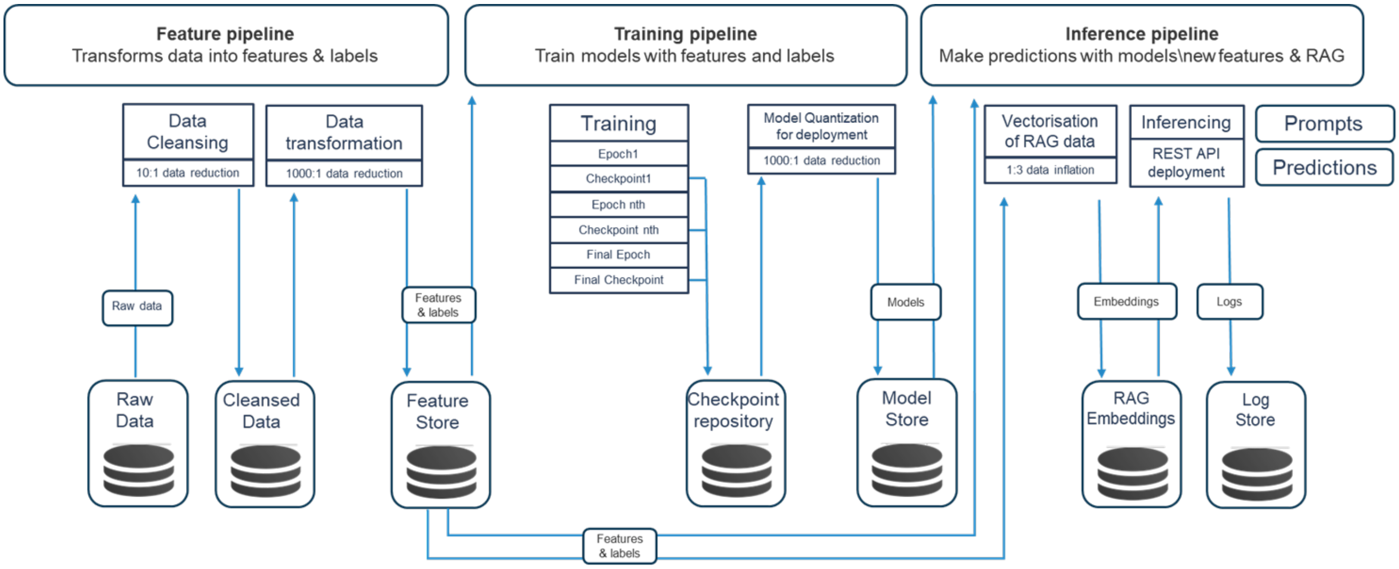


Figure 5: High-level view of a data processing pipeline

This paper is focussed on the training and inference data pipeline workload activities as the HGX and OVX compute platforms are specifically designed for these workloads.

Data ingestion stage

As data is created it needs to be stored, prepared and transformed ready to be usable by model training frameworks. A suitable data storage platform must be able to deal with a multitude of data object types and sizes across a variety of read and write patterns. Each dataset ingestion activity may differ in its data IO impact on the storage platform, so intelligent management and access to data is required.

Feature pipeline

The data stored in raw format is unlikely to be immediately digestible by a model training programme as it needs to be cleansed and transformed, like “Big Data’s” ETL (Extract, Transform, Load) processes.

Data cleaning techniques help ensure that data is accurate and consistent across different datasets. Data transformations prepare data for analysis - reshaping and computing new dimensions and metrics. The type of data cleansing and transformations required depend on the raw data format and how that differs from the format required to be ingested into the machine learning engine.

Training pipeline

Once the data has been cleaned and prepared, features selected and training labels applied, the data can be ingested into the training framework and applied to the machine learning engine. The size of the training dataset is an important consideration here. Typically, we are looking at 10PBs of training data for larger LLMs and as the size of the processing cluster increases (more GPUs added), so does the rate at which data is ingested.

The IO workload for this activity is typically a large data object sequential read. As the number of GPUs in the training supercomputer increases the faster data needs to be read into GPU memory. Typical file formats used for data ingestion are large files that have been constructed to allow data to be easily distributed across multiple GPUs. The storage system must keep the GPU memory full to ensure no interruptions to GPU processing time.

Training checkpoints are scheduled periodic snapshots of the model training process in flight and provides a method of recovering a training process to a point in time. In the event of a training job being interrupted, rather than restart the training process from the beginning, training can restart from the last known good checkpoint. Checkpoint file sizes are typically large, dictated by the size of the model being trained in regard to the number of parameters.

Inferencing pipeline

Once we have a trained model that can be deployed in production, another set of activities start up that put different pressures on the storage platform. An AI model’s responses can only be based on that data that’s been used for training. In fact, an AI’s responses may be out of data as soon as it’s deployed depending on the nature of the queries submitted. One way to update a model’s references is to retrain on additional datasets. But training is a lengthy and costly exercise, and more time will be spent on training than delivering value. To increase the breadth of knowledge and keep an AI up to date, external references sources are required.

The most common method of providing external data sources to Ais during inferencing is known as Retrieval-Augmented Generation (RAG) and is a knowledge enhancement technique which enables domain specific external datasets to provide information to the trained model. An example is converting a pdf user manual into vector data so that it can be queried and used to enhance the response from the AI model.

RAG has some limitations, response time for lengthy queries for example, as the user query needs to be converted from KV tokens into a vector space and performing similarity searches against data already stored in the vector database. Once matches are found, the vector data needs to be converted back to tokens which the AI uses for making decisions. The KV token data is typically stored in a Key Value Cache (KV cache) in the GPU memory and provides for high-speed processing of response to queries.

A new methodology for managing this external data referencing is called Cache Augmented Generation (CAG) which delivers a greater response time for lengthy contextual queries. This process avoids the vector database completely, by using the inferencing capabilities of the AI model to convert external raw data (pdf files etc..) directly into KV token format and stored in a KV Cache file which is stored on the storage platform and moved into GPU memory as required.

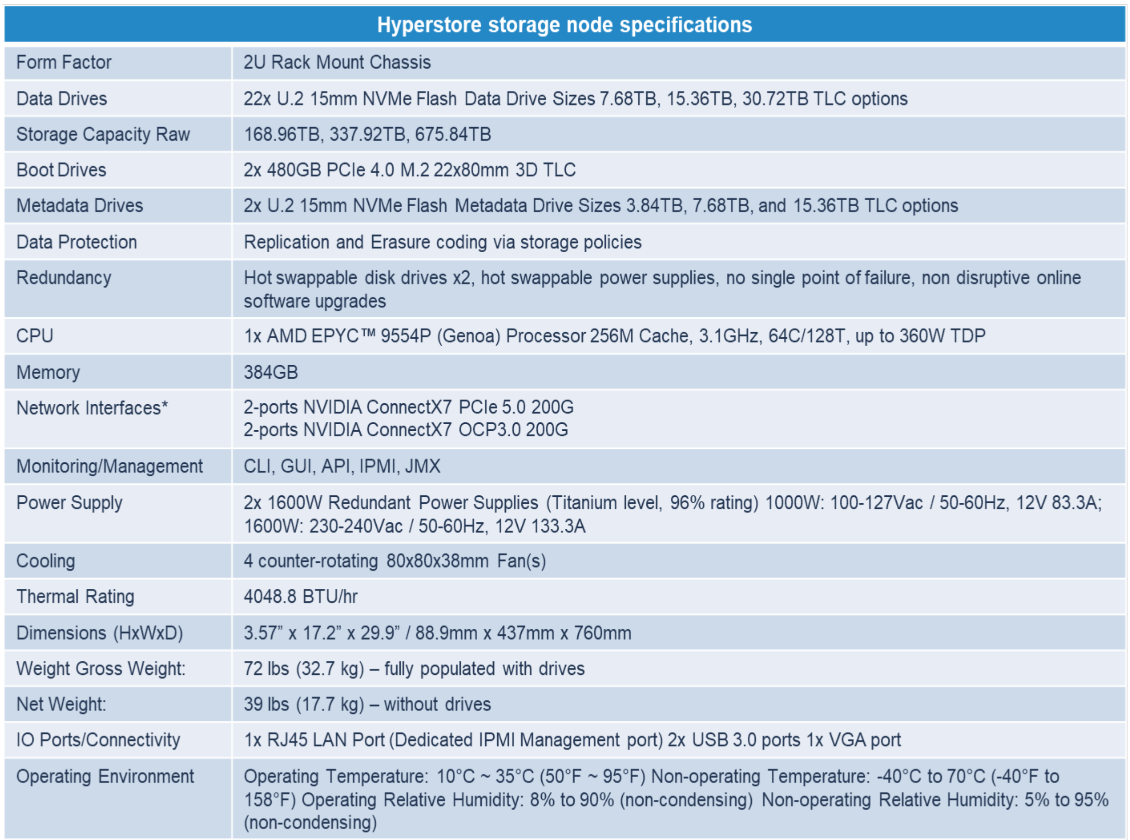
In both use cases data is stored in large data files of hundreds of MBs to GBs in size, with those files created and read multiple times.

Solution Components

HyperStore Storage Nodes - Supermicro AS-2115HS-TNR

Figure6: The Supermicro "Storage Server AS-2115HS-TNR"

The HyperStore nodes are Supermicro Hyper A+ Servers (model no. AS-2115HS-TNR) pre-integrated with Cloudian HyperStore software. This hardware platform has been designed specifically for AI and machine learning workloads where deterministic access times and lower latency is critical. Appliances are available with different options for metadata drives, data drives, networking connectivity, and support packages including options for on-site service and support. A full "Bill of Materials" is specified for this server complete with Supermicro part numbers in Appendix 1. Specifications and options for the storage nodes are shown in Table 1.



Hyperstore storage node specifications	
Form Factor	2U Rack Mount Chassis
Data Drives	22x U.2 15mm NVMe Flash Data Drive Sizes 7.68TB, 15.36TB, 30.72TB TLC options
Storage Capacity Raw	168.96TB, 337.92TB, 675.84TB
Boot Drives	2x 480GB PCIe 4.0 M.2 22x80mm 3D TLC
Metadata Drives	2x U.2 15mm NVMe Flash Metadata Drive Sizes 3.84TB, 7.68TB, and 15.36TB TLC options
Data Protection	Replication and Erasure coding via storage policies
Redundancy	Hot swappable disk drives x2, hot swappable power supplies, no single point of failure, non disruptive online software upgrades
CPU	1x AMD EPYC™ 9554P (Genoa) Processor 256M Cache, 3.1GHz, 64C/128T, up to 360W TDP
Memory	384GB
Network Interfaces*	2-ports NVIDIA ConnectX7 PCIe 5.0 200G 2-ports NVIDIA ConnectX7 OCP3.0 200G
Monitoring/Management	CLI, GUI, API, IPMI, JMX
Power Supply	2x 1600W Redundant Power Supplies (Titanium level, 96% rating) 1000W: 100-127Vac / 50-60Hz, 12V 83.3A; 1600W: 230-240Vac / 50-60Hz, 12V 133.3A
Cooling	4 counter-rotating 80x80x38mm Fan(s)
Thermal Rating	4048.8 BTU/hr
Dimensions (HxWxD)	3.57" x 17.2" x 29.9" / 88.9mm x 437mm x 760mm
Weight Gross Weight:	72 lbs (32.7 kg) – fully populated with drives
Net Weight:	39 lbs (17.7 kg) – without drives
IO Ports/Connectivity	1x RJ45 LAN Port (Dedicated IPMI Management port) 2x USB 3.0 ports 1x VGA port
Operating Environment	Operating Temperature: 10°C ~ 35°C (50°F ~ 95°F) Non-operating Temperature: -40°C to 70°C (-40°F to 158°F) Operating Relative Humidity: 8% to 90% (non-condensing) Non-operating Relative Humidity: 5% to 95% (non-condensing)

Table 1. HyperStore storage nodes - Supermicro AS-2115SH specifications and options.

Micron 6500 ION NVMe SSD

The HyperStore storage node contains 22 Micron 6500 ION SSDs to deliver the performance and capacity demands required by AI and ML workloads, whilst also trying to manage environmental impact. The Micron 6500 ION SSD addresses all these problems head-on with best-in-class performance that promotes sustainability and eclipses the QLC SSD with superior value.

Deployed with 30.72TB capacity and dense form factors, the high-speed TLC SSD provides greater write endurance than QLC at an equivalent price. QLC SSDs are suited to longer term archiving of data and as the technology capacity grows then it becomes relevant in a cost per GB discussion, but workload types will always be limited on QLC.

Figure 7: Micron 6500 ION NvMe SSD

Cloudian HyperStore

Cloudian Hyperstore is a software defined data platform providing data storage and management services for the largest datasets with scalability to exabyte capacity levels. HyperStore uses "Object" based data management for data storage, protection and retrieval with the S3 API as the primary data access interface for users and applications with file protocols (NFS & SMC) also supported providing bi-modal access to singular data sets.

Built on cloud scale design and technologies, HyperStore clusters can be deployed in a multitude of architecture designs such as private cloud, globally distributed clusters, edge to cloud, data backbone as a Service for other services.

Cloudian HyperStore architecture

Cloudian HyperStore virtualises the physical storage resources within the cluster by presenting a virtual container called an "S3 bucket" to client users\systems to store data files regardless of type and\or format.

The HyperStore engine takes care of data placement across all cluster resources, guided by user defined storage policies. Storage policies provide object level granularity for controlling where and how data will be stored. Any information stored in the objects metadata can be used as input to make decisions on how\where data is stored within the cluster.

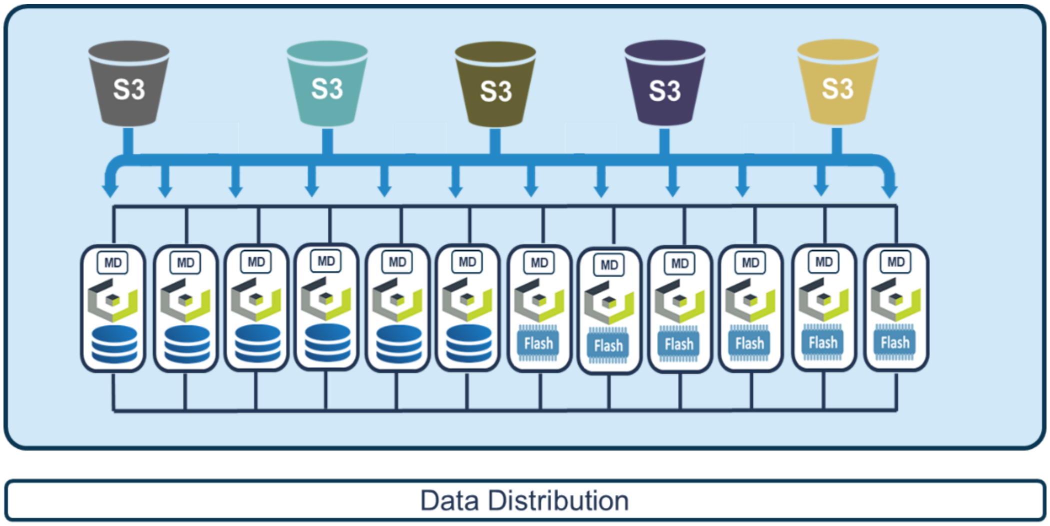


Figure 8: HyperStore scale out clustered architecture

Performance

Traditionally object storage has always been seen as a highly scalable capacity archive storage tier for data that is less frequently accessed and maybe only kept for "just in case" scenarios and certainly not primary application storage for high performance workloads. For larger data files written to storage media in a sequential manner, object storage has always had a competitive edge on traditional scale up file and block storage systems, due to its distributed parallel architecture. Block and file solutions including NFS are typically single threaded operations from the storage initiator to storage target.

Parallelisation

HyperStore is designed to simplify data storage and access at scale using a scale out cluster architecture design. Scale out designs are typical of high-performance compute (HPC) platforms as data requests can be parallelised across all hardware resources within the cluster and enables data to be transferred in the fraction of the time it takes for a single sequential data transfer to a limited number of resources. HyperStore uses the S3 API to break up large files into smaller parts and distribute to all the available storage resources in the storage cluster in a parallel, multi-threaded distribution, for HyperStore, every single thread is automatically converted to multi-threaded requests. Similarly, for reads, data is read from all resources used to store the multiple part writes. This design eliminates bottlenecks and allows performance to grow as you add more cluster nodes.

Hyperstore can parallelise data threads both at the client end with multi-part upload and as the data hits the storage cluster, by balancing data requests across all cluster nodes and drives automatically.

Metadata

Object storage uses metadata to manage user access and location placement for stored data. At time of upload the data object will have a metadata entry created in the database. Automatically various system attributes will be assigned to the data object, and optionally additional custom metadata can be added either by user or automated updates.

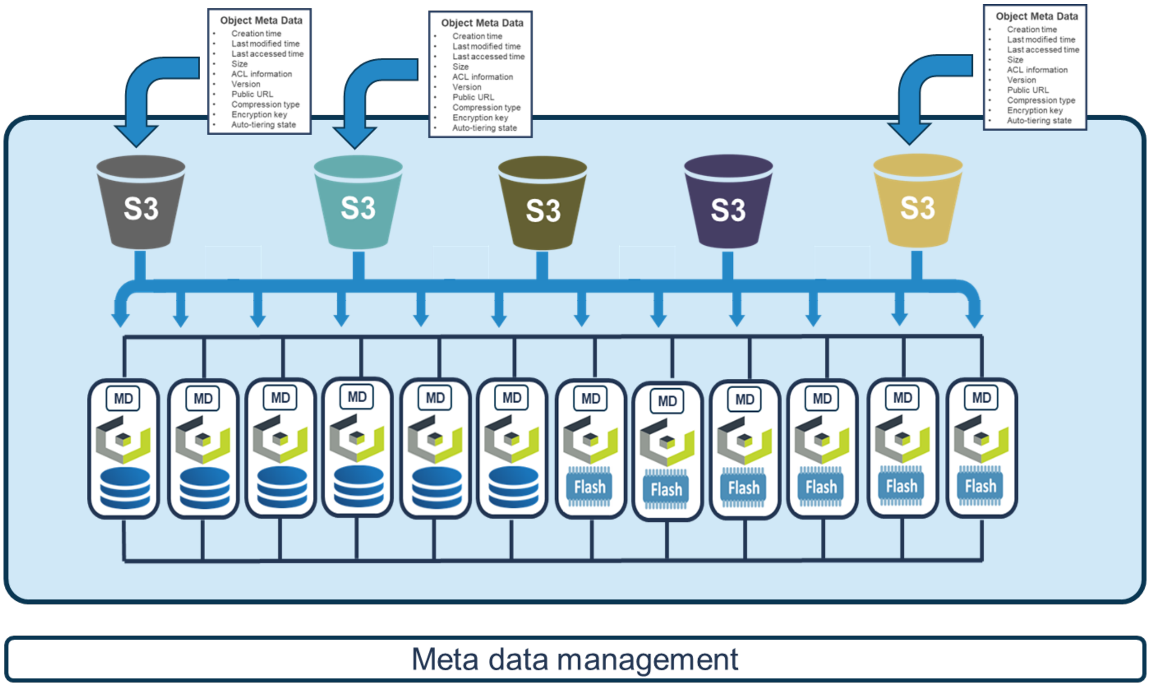


Figure 9: How metadata is stored to reference actual data objects

As an example, the metadata for a digital photo image, metadata could be:

System generated metadata

- Unique identifier
- GPS location
- Image size
- Created by
- Resolution
- Data taken
- Confidential
- Expiration

Custom metadata describing photo features

- Category of data
- Cat
- Dog
- Person
- Location
- Clothing
- ….

The metadata can contain far more information relating to the data object type than a data file or block. Everything stored is given a unique metadata identifier and every piece of data has the metadata and file combined into a single object.

An object can be any type of file and metadata provides searchable capability across different file types and data sources, providing a unique opportunity to combine data searches across multiple datasets and formats.

Metadata turns unstructured data into a structure with a common searchable format across multiple types. The new frontier is to provide model inference to enrich existing data with more descriptive metadata that can be added on ingestion. Metadata is a lot easier to move around a network than actual data and can be extracted from any data type.

Multi-tenancy

HyperStore is designed for cloud-like multi-tenancy, providing a consolidated platform for ease of management and economies of scale that allows multiple users to securely share a single environment, with unique namespaces and delegated management functions. Offering a range of service level configurations is also critical as not all tenants want the same thing and of course not all data is created equal.

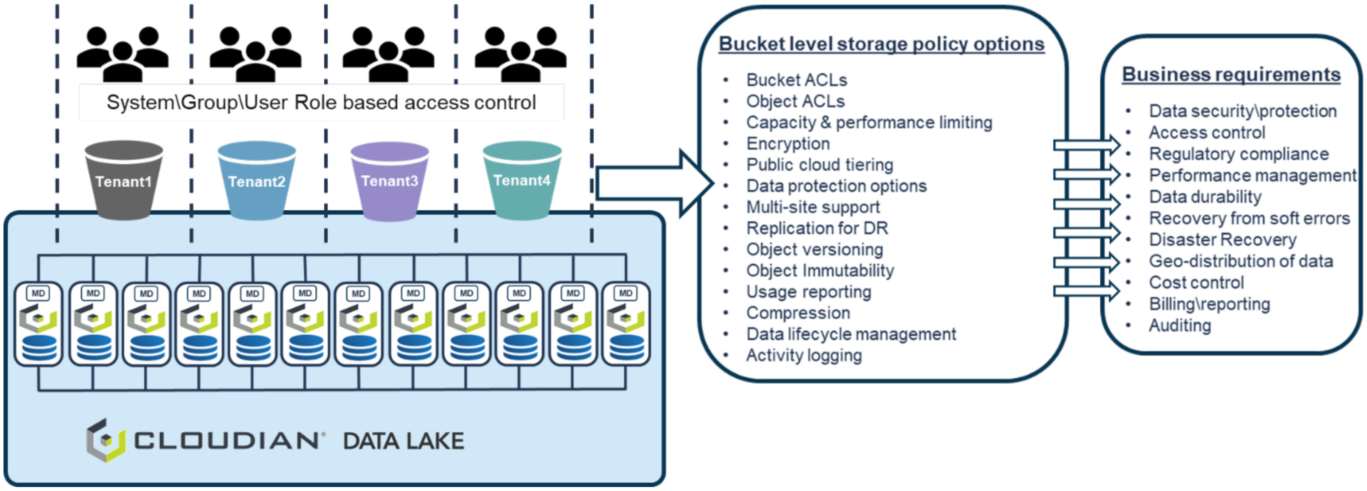


Figure 10: Logical multi-tenancy – shared everything architecture

Users, whether they're human or machines, are managed and authenticated to the storage system through the usual user management tools such as LDAP, Active Directory and IAMs to gain access to relevant storage resource.
S3 buckets are the logical data container presented to users for the storage of data objects. Storage policies are applied at the bucket level which determines where and how the data objects in this bucket are stored.

Storage policies can be created and controlled either by administrators and\or users depending on deployment requirements and should be configured to match the business needs of the datasets stored in that bucket. Examples of configurable options within storage policies are shown in diagram above.

HyperStore supports KMIP standardized key management as a Server-Side Encryption (SSE) with AES-256 and FIPS-validated crypto-graphic modules. SSL encryption ensures data confidentiality for data in transit (HTTPS). And with S3-compatible ACLs, SAML, MFA, IAM, and secure trust policies, system administrators can better manage access to buckets and objects.

Physical multi tenancy

HyperStore can also provide physical separation for tenants that require guaranteed performance vs sharing a larger pool of resources that may have limiting levels applied. Hardware separation is particularly requested to support workloads such as AI model training with high bandwidth requirements needed for data ingest and for checkpointing schedules. The below diagram shows three different deployments.

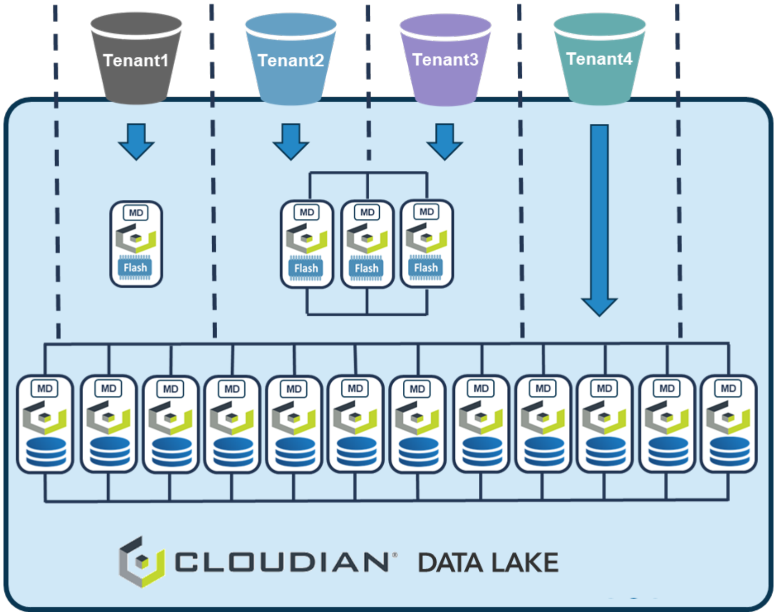


Figure 11: Physical multi-tenancy – dedicated HW resources

Tenant 1 is configured with a dedicated HyperStore Flash instance to serve as a high-speed cache using TLC NvMe solid state storage devices backed up by a shared hard disk or QLC NvMe based storage pool, providing longer term retention at lower cost. Data is preloaded from 2nd tier storage media to 1st tier ahead of model training runs.

Checkpointing is immediately staged to the fastest storage to allow GPUs to restart training as soon as possible. Data is then moved to the 2nd tier for cost optimization based on specific requirements for the recovery processes.

Tenants 2 & 3 each have a secure tenancy but are sharing high speed physical resources as well as tier 2 resources. This configuration may be more suitable for data cleansing and transformation workloads which are based on processing changes to the data objects as they are ingested in real time. Different ingest patterns could be analysed to understand peak performance requirements that could be shared on the same resources.

Tenant 4 does not really have any major performance requirements and is satisfied with writing directly into the safe virtual bucket distributed across the 2nd tier.

Cloudian HyperIQ

Cloudian HyperIQ provides a centralized management and monitoring software with data and actionable insights to monitor and manage the HyperStore data lake infrastructure by collecting monitoring and operational data from logs, system metrics, and events, providing a unified view of user and system resource activity.

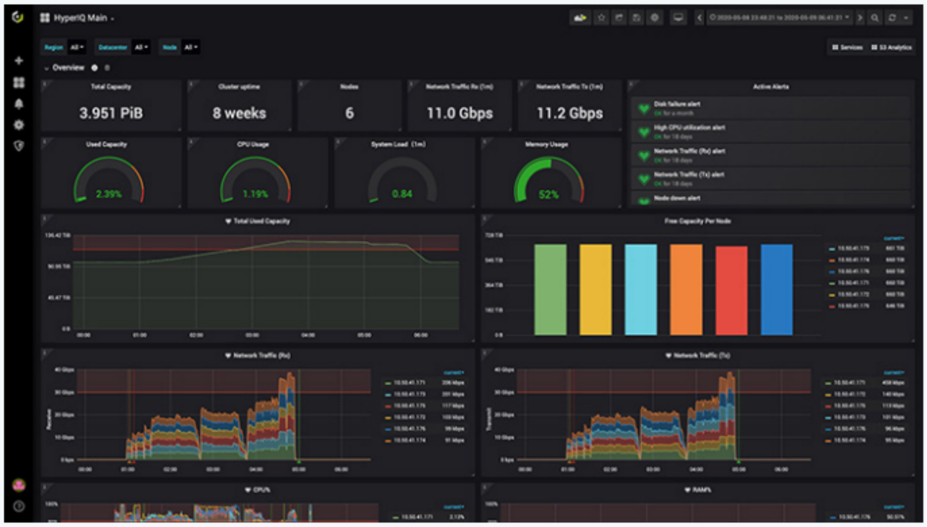


Figure 12: Cloudian HyperIQ – Infrastructure insights and monitoring

HyperIQ provides a unified view of operational health with an end-to-end view of performance, from client to media which allows for optimized resource utilization. Performance monitoring is especially important for AI and ML workloads, as it is crucial to keep all GPU's fed with data to minimise GPU idle time. High-speed read responses are required for these processes.

Similarly, with training checkpoints required to save training models at various intervals the checkpoint must be completed as quickly as possible as the training needs to pause to ensure data consistency between all the GPUS used across the model training. High-speed write responses are required for these processes. All aspects of the cluster are observable to quickly identify and understand issues potentially before they happen to.

Network Architecture

NVIDIA Spectrum-X networking

The NVIDIA Spectrum-X Ethernet platform is designed specifically to improve the performance and efficiency of Ethernet-based AI deployments, achieving 1.6X better networking performance, along with consistent, predictable performance with **large scale and multi-tenant** environments. Spectrum-X is built on tight network integration of the NVIDIA Spectrum-4 switches, LinkX 400G cables and NVIDIA BlueField-3 SuperNICs and DPUs.

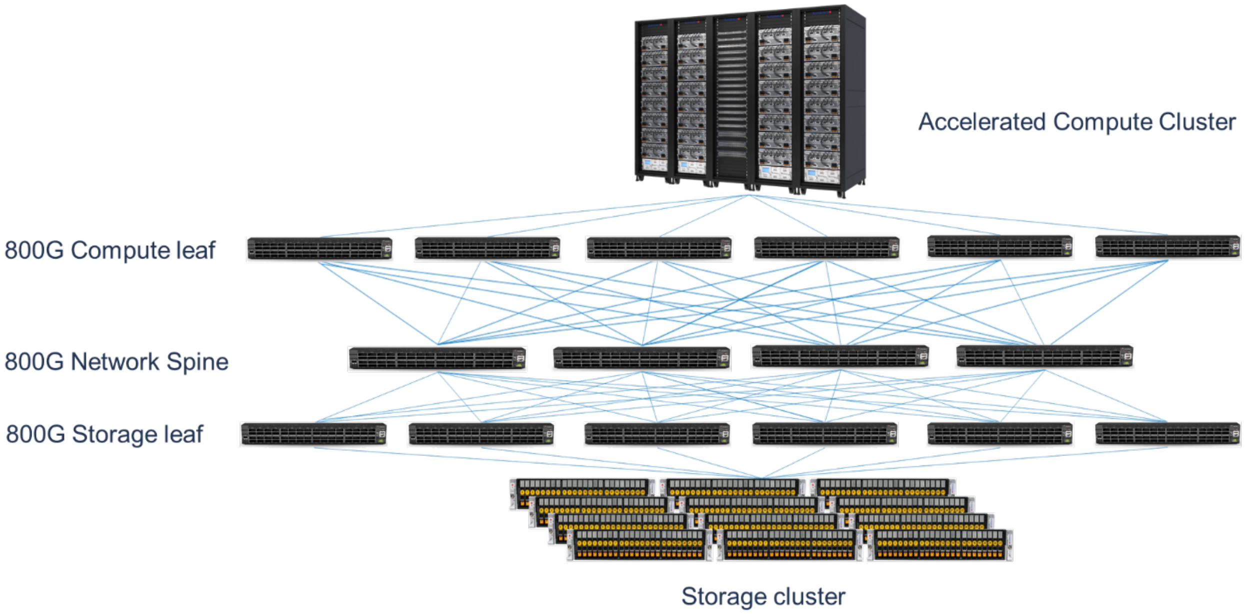


Figure 13: Spectrum-X networking fabric design

AI workloads require a low-latency network with high, effective bandwidth that can run many jobs simultaneously and isolate tenant usage patterns to prevent interference amongst tenants running on the same compute and network infrastructure. Spectrum-X leverages remote direct-memory access (RDMA) over Converged Ethernet (RoCE) extensions to deliver both high-performance networking and cloud multi-tenancy,

End-to-End Adaptive RDMA Routing with Lossless Network

- BlueField-3 sends data into the switch network
- Spectrum-4 spreads the data packets across all available routes
- BlueField-3 ensures in-order data delivery
- Increase from typical 60% to 95% effective bandwidth

Noise Isolation with Programable Congestion Control

- Diverse workloads can impact each other's performance
- Spectrum-X detects congestion spots in real time
- Programmable congestion control meters the data flow
- Results in performance isolation across workloads

The BlueField-3 SuperNIC is a new class of network accelerator that's purpose-built for hyperscale AI workloads. Designed for network-intensive, massively parallel computing, the BlueField-3 SuperNIC provides best-in-class remote direct-memory access over converged Ethernet (RoCE) network connectivity between GPU servers at up to 400Gb/s, optimizing peak AI workload efficiency. For modern AI clouds, the BlueField-3 SuperNIC enables secure multi-tenancy while ensuring deterministic performance and performance isolation between tenant jobs.

Figure 14: NVIDIA Bluefield 3 DPU and SuperNIC

The NVIDIA Spectrum SN5600 switch can be deployed as leaf, spine, and super-spine switch roles. The SN5600 provides 64 ports of 800GbE in a 2U form factor offering diverse connectivity combinations between 1 to 800GbE and delivers a total throughput of 51.2Tb/s.

Figure 15: NVIDIA Spectrum-X SN5600 switch

Storage node connectivity

The storage node should be configured with two dual port NICs providing a total of four 200Gb/s capable ports. When building a resilient storage network, best practise is to configure multiple NICs across multiple switches to provide for NIC and switch level failure. Each NIC should have one port connected to a switch that is different from its other port.

The MMA4Z00 transceiver installed in the SN5600 switch converts the single 800Gb/s port into two 400Gb/s ports. Using split MPO-12/APC to MPO-12/APC cables each 400Gb/s switch port can connect to two 200Gb/s end points. These should be distributed across the two separate NICs in the storage node. This is repeated for both switches.

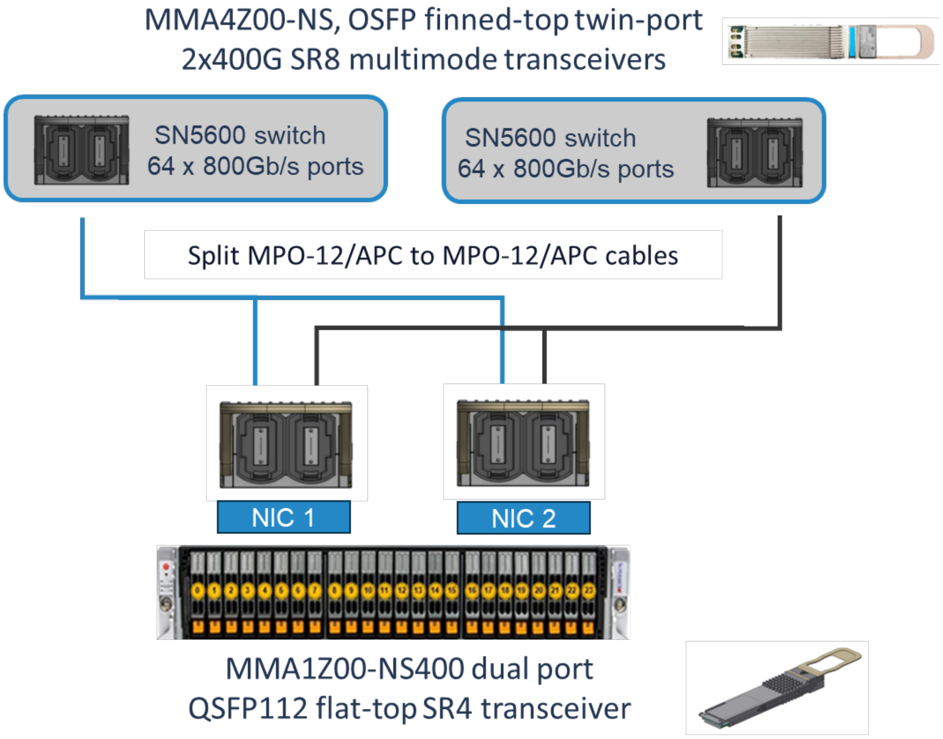

MMA4Z00-NS, OSFP finned-top twin-port
2x400G SR8 multimode transceivers
SN5600 switch
64 x 800Gb/s ports
SN5600 switch
64 x 800Gb/s ports
Split MPO-12/APC to MPO-12/APC cables
NIC 1
NIC 2
MMA1Z00-NS400 dual port
QSFP112 flat-top SR4 transceiver

Figure 16: Storage node network connectivity

The HGX cluster reference design includes several discrete networks, which can be virtually or physically separated depending on switching infrastructure and business requirements. This network is designed to meet the high-throughput, low-latency, and scalability requirements of AI workloads running on HGX clusters.

1. **Front end storage cluster network** – Connects ports on the HyperStore nodes to the compute clients over 2 x 200Gbps ethernet ports via the NVIDIA Spectrum SN5600 switch. This network provides connectivity to satisfy the client computers IO requests from the storage platform. Shown as dark blue connections.

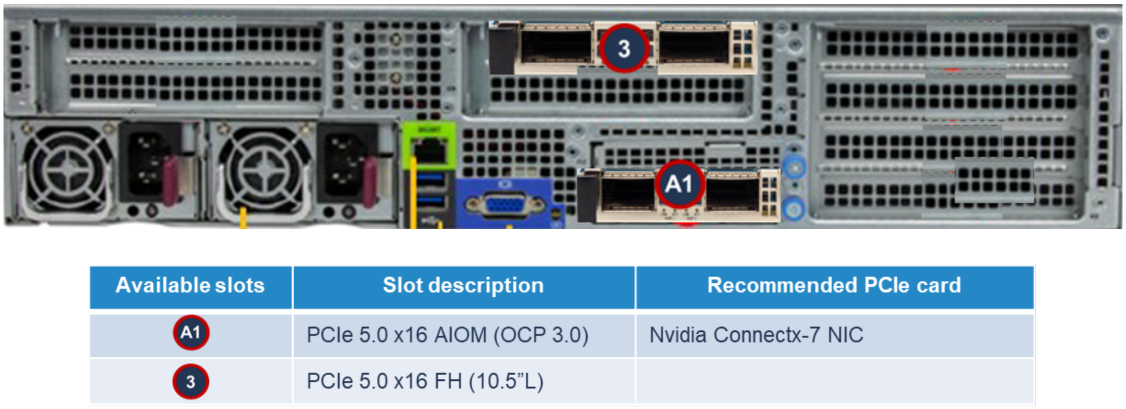
2. **Back-end storage cluster network** - Connects ports on the HyperStore nodes to other HyperStore nodes over 2 x 200Gbps ethernet ports via the NVIDIA Spectrum SN5600 switch. This network is used as the inter cluster communication backbone transmitting metadata and data for multi node data distribution and data protection/DR purposes. Shown as green connections.

3. **Management network** – Outside of the data flow, the management network provides a single network to gain access to the various hardware devices for system management and monitoring. Shown as light blue connections.

HyperStore node network connectivity

It is advised that this storage node configuration with 24 drives reduces the availability of PCI slots can be used for expansion as the other slots are assigned to NvMe drives

To maximise the bandwidth Cloudian recommends 2 x NVIDIA ConnectX-7, BF3 DPU or BF3 SuperNICs with dual ports which require 16x lane PCIe Gen 5.0 slots. Slots 3 and A1 remain to be used and both support 16x lanes. Note figure 11, slot 3 is standard PCI3 gen 5.0. but slot A1 is an OCP 3.0 form factor, so be careful when ordering as one of each is required per storage node.



Available slots	Slot description	Recommended PCIe card
A1	PCIe 5.0 x16 AIOM (OCP 3.0)	Nvidia Connectx-7 NIC
3	PCIe 5.0 x16 FH (10.5"L)	

Figure 17: Recommended HyperStore node network port connections

All four ConnectX-7 ports on each storage node should be connected to the switch layer with one port per card connected to each switch for network fault tolerance as shown on the network diagram in figure 10. As well, the 1 GB/s management\BMC port should be connected to the out-of-band management network.

Network configuration details

The network setup is a standard installation. Full network configuration details are beyond the focus of this document. Two key points to note,

1. NICs must have the RDMA driver installed
2. Jumbo frames need to be enabled (MTU=4200) on the NIC bonds and the bond VLAN tagged interfaces

Testing environment and methodology

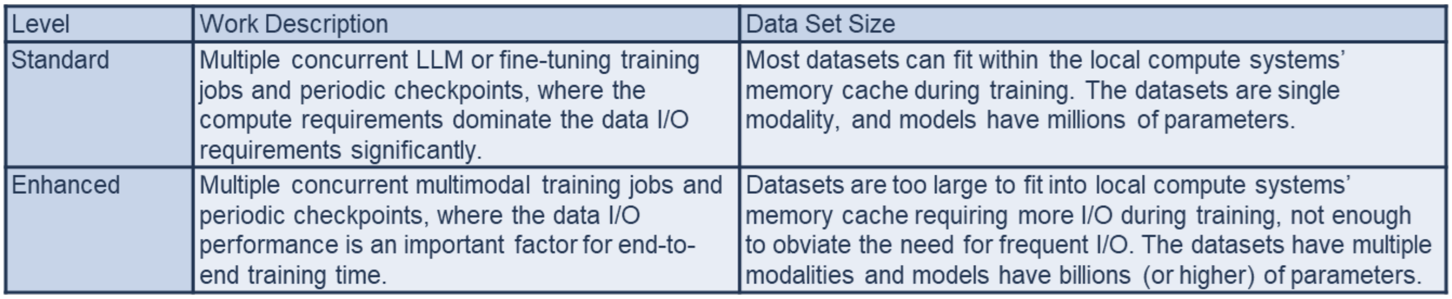
The performance benchmarking tool used to validate this reference architecture is called GOSBench and will be used in this analysis to generate a range of differing workloads of various object sizes, read\write patterns and concurrent transactions to simulate different AI workloads.

GOSBench originated from an initial project from Intel in 2013 called Cos bench which was developed for object storage performance benchmarking tool and originally written in Java V6.

The GOSBench project (https://github.com/mulbc/gosbench) was started in July 2019 and is written in Golang with active development on GitHub. Open source, Golang, thus very portable with just 2 binaries, go server and gosdriver. This is a vendor independent tool using the AWS S3 libraries for workload generation.

For specific testing with S3 over RDMA, Cloudian created a dynamic loadable library to integrate with NVIDIA Cuda to allow the following:

- Creates/closes cuobject client
- Allocate and register RDMA memory buffers
- Deallocate and unregister RDMA memory buffers
- Reads and Writes data from RDMA memory buffers
- Handles putObject and getObject calls



Level	Work Description	Data Set Size
Standard	Multiple concurrent LLM or fine-tuning training jobs and periodic checkpoints, where the compute requirements dominate the data I/O requirements significantly.	Most datasets can fit within the local compute systems' memory cache during training. The datasets are single modality, and models have millions of parameters.
Enhanced	Multiple concurrent multimodal training jobs and periodic checkpoints, where the data I/O performance is an important factor for end-to-end training time.	Datasets are too large to fit into local compute systems' memory cache requiring more I/O during training, not enough to obviate the need for frequent I/O. The datasets have multiple modalities and models have billions (or higher) of parameters.

- Use Golang Cgo to bind the dll to the GOSBench Driver. This allows golang code to call C/C++ functions directly
- Added an option in the S3 config to select between GPU memory or System memory for RDMA.

GOSBench gosdriver was configured to run on each test server to generate different workloads to understand behaviour of the HyperStore storage platform under sustained load.

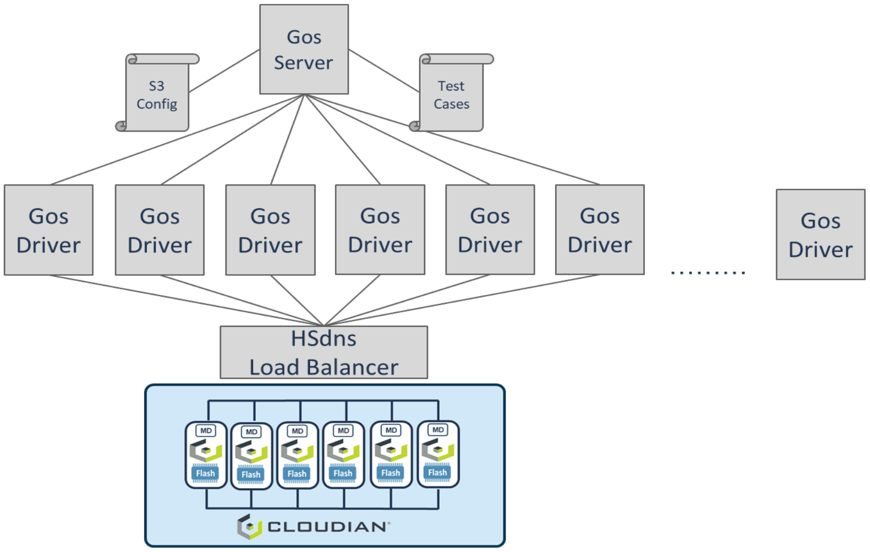


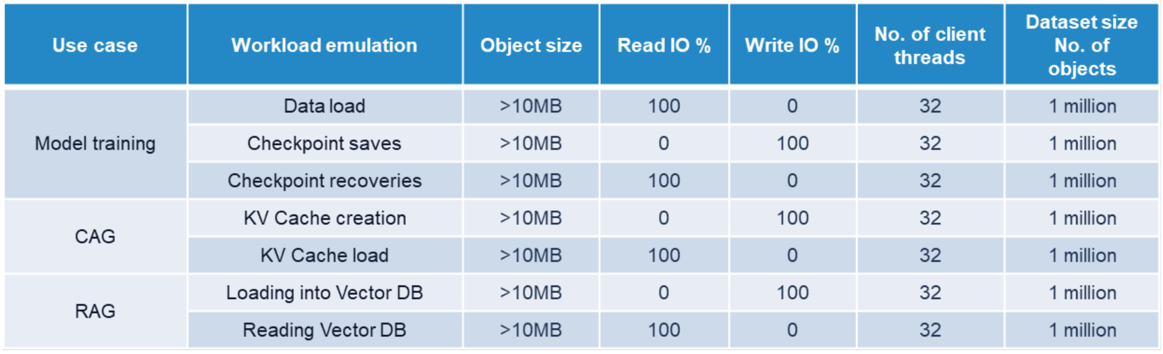
Figure 18: Cloudian HyperStore performance testing setup

GOSBench can be configured with a range of parameters that allow it to be used in emulating a variety of data IO patterns that can be matched to real world examples.

GOSBench parameters that can be adjusted include:

- Object size
- No. of client threads
- No. of clients
- Read\Write ratios
- No. of objects used in a test
- Duration of tests

For this testing, the parameters were configured to emulate seven specific workloads, although all seven use cases share only two common IO profiles.



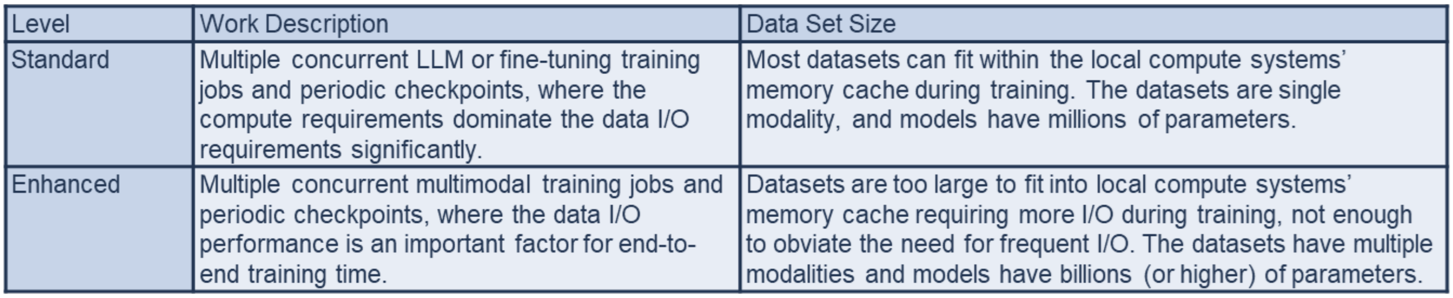
Use case	Workload emulation	Object size	Read IO %	Write IO %	No. of client threads	Dataset size No. of objects
Model training	Data load	>10MB	100	0	32	1 million
	Checkpoint saves	>10MB	0	100	32	1 million
	Checkpoint recoveries	>10MB	100	0	32	1 million
CAG	KV Cache creation	>10MB	0	100	32	1 million
	KV Cache load	>10MB	100	0	32	1 million
RAG	Loading into Vector DB	>10MB	0	100	32	1 million
	Reading Vector DB	>10MB	100	0	32	1 million

Table 2: Performance test configurations

The results and resulting proposed architectures are based upon the results of these tests configured across a range of clients and storage nodes within the data cluster.

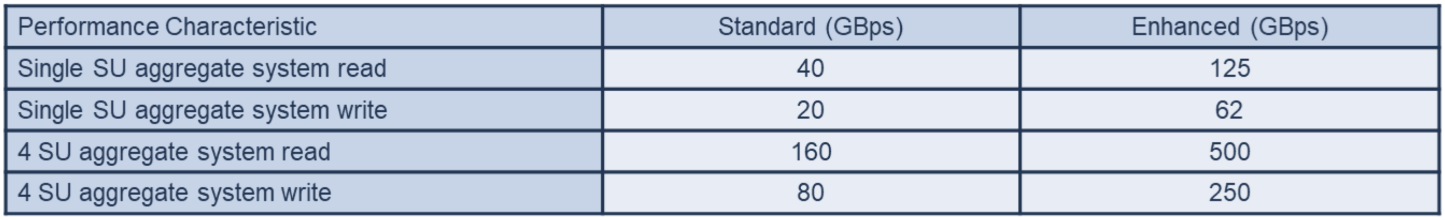
Validated Solution summary

The storage system performance required to maximize training performance can vary depending on the type and size of model being trained and the size of the training dataset. NVIDIA provide guidance for sizing storage systems to support training workloads running on NVIDIA DGX compute platforms. This can be found at https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-b200/latest/storage-architecture.html. Tables 3 and 4 below have been taken from this guide and will be used as the basis for sizing the storage platform to support accelerated compute workloads running on Supermicro HGX and OVX servers.



Level	Work Description	Data Set Size
Standard	Multiple concurrent LLM or fine-tuning training jobs and periodic checkpoints, where the compute requirements dominate the data I/O requirements significantly.	Most datasets can fit within the local compute systems' memory cache during training. The datasets are single modality, and models have millions of parameters.
Enhanced	Multiple concurrent multimodal training jobs and periodic checkpoints, where the data I/O performance is an important factor for end-to-end training time.	Datasets are too large to fit into local compute systems' memory cache requiring more I/O during training, not enough to obviate the need for frequent I/O. The datasets have multiple modalities and models have billions (or higher) of parameters.

Table 3: NVIDIA workload categorisation for storage IO impact



Performance Characteristic	Standard (GBps)	Enhanced (GBps)
Single SU aggregate system read	40	125
Single SU aggregate system write	20	62
4 SU aggregate system read	160	500
4 SU aggregate system write	80	250

Table 4: NVIDIA guidelines of performance required from data storage systems as GPU clusters are scaled up.

The aim of this reference architecture is to demonstrate how a HyperStore design can be scaled to meet the "enhanced" performance recommendations for datasets that are too large to fit into cache, have massive first epoch I/O requirements and workflows that only read the dataset once. This is typical of uses cases such as training with 1080p, 4K, or uncompressed images, offline inference and ETL workloads.

The below results show recommended Cloudian HyperStore cluster configurations as a HGX GPU cluster is scaled from a single NVIDIA Super pod Scalable Unit (SU) equivalent (32 GPU nodes with 256 GPUs) to sixty-four SUs (2047 GPU nodes with 16384 GPUs).

# of NVIDIA DGX SuperPOD SU size	# of HGXs	# of GPUs	Target Write perf (GB/s)	Target Read perf (GB/s)	# of storage nodes	Rack units	Raw capacity (PB)	Usable capacity (PB)	Write perf (GB/s)	Read perf (GB/s)
1	**32**	**256**	**62**	**125**	**9**	18	**6**	4.6	**90**	**180**
2	**64**	**512**	**125**	**250**	**18**	36	**12**	9.1	**180**	**360**
4	**128**	**1024**	**250**	**500**	**27**	54	**18**	13.7	**270**	**540**
8	**256**	**2048**	**500**	**1000**	**54**	108	**36**	27.3	**540**	**1080**
16	**512**	**4096**	**1000**	**2000**	**108**	216	**73**	54.7	**1080**	**2160**
32	**1024**	**8192**	**2000**	**4000**	**216**	432	**146**	109.4	**2160**	**4320**
64	**2047**	**16384**	**4000**	**8000**	**432**	864	**292**	218.7	**4320**	**8640**

IO performance results based upon <10MB data objects.
32 workers per client
Data protection using Erasure Coding 9+3

Table 5: Performance test results

This validation confirms functional integration, and optimal performance out-of-the-box for providing data storage and management services to accelerated compute cluster configurations.

- GPU-accelerated applications can engage the full capabilities of the data infrastructure and the HGX systems.
- Performance is distributed evenly across all the HGX systems in the cluster
- Scales linearly as more HGX systems are engaged.
- Scales linearly as more storage nodes are added to the cluster

Summary - All you need is …Object

The explosion in interest and innovation in the Artificial Intelligence industry is startling for many reasons, but a more prosaic impact, compared to robot training, AGI etc., is on the IT infrastructure that is needed to power such advancements. Especially in the data storage and management solutions area, where the data access performance demands can only be met by High Performance Computing (HPC) storage architecture where parallel data processing is done at a scale like never before and by more organisations that ever before.

With the advances Cloudian have made with developing a high-performance object storage platform, organisations can now look to develop, enrich and make available their data sets like never before without a concern on being able to support all high-performance workloads. Building data pipelines around the storage platform is the next step where a data object's lifecycle can be aligned with the data processing pipelines and achieving more business value from combining different datasets and looking at problems from different angles.

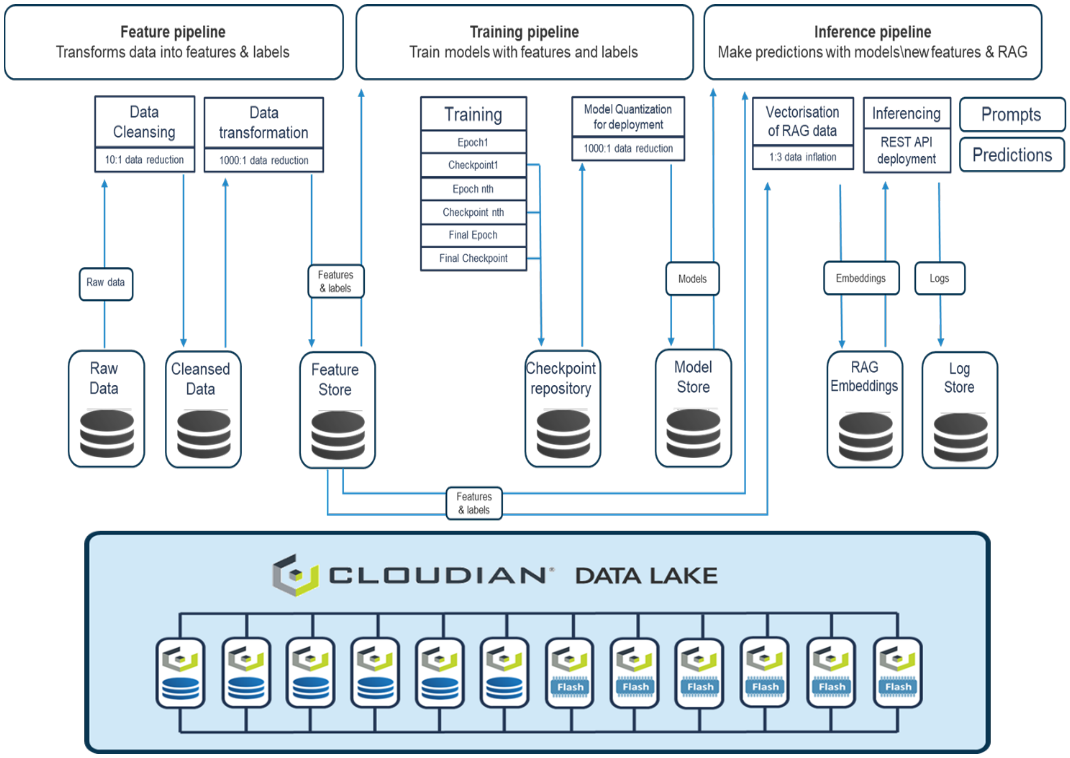
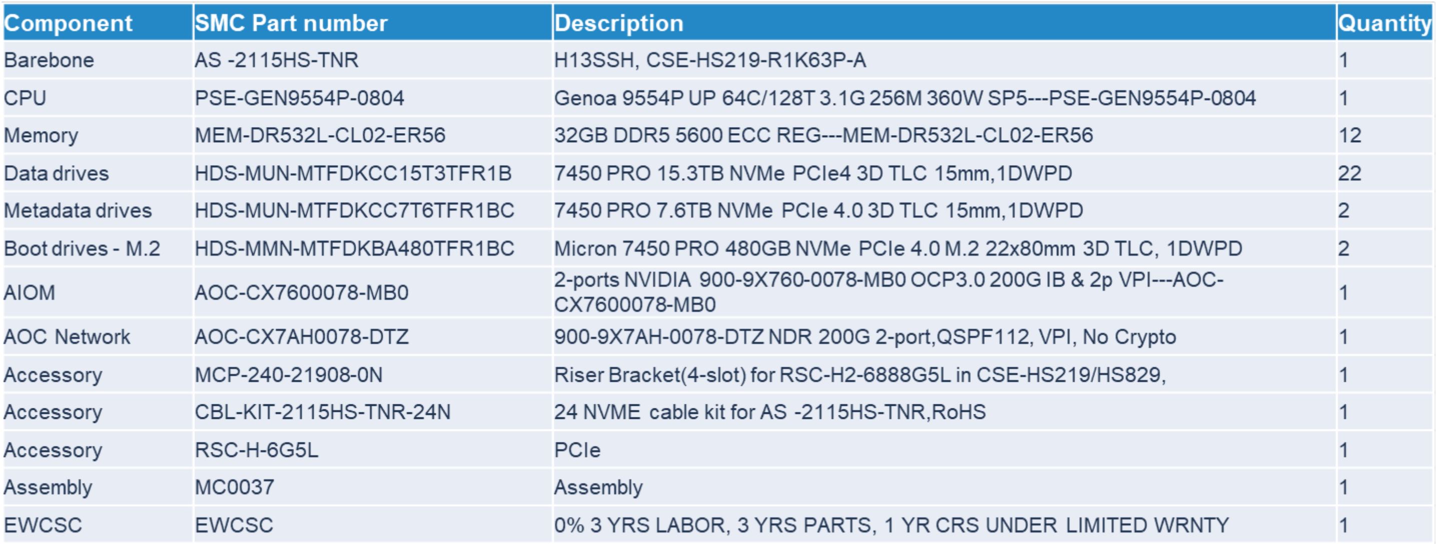


Figure 19: Cloudian HyperStore is the only storage platform needed for all AI and ML workloads

But even more important than being able to access data rapidly, is the quality AND quantity of data that is required to push AI learnings to the limits. Data must be gathered and prepared carefully before submitting to model training, to ensure accurate, unbiased, truthful data is input. And the more examples of data features that are collected and submitted brings the power of AI up immeasurably.

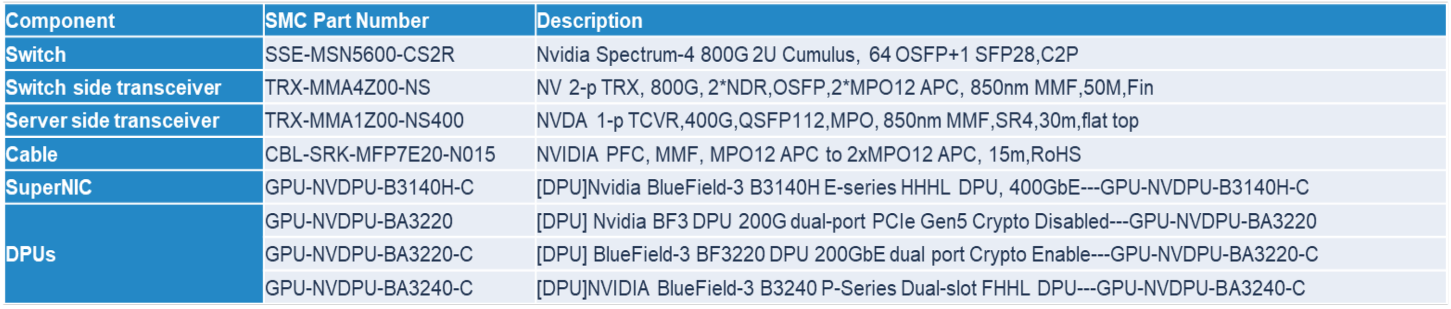
Appendix

Supermicro storage node BOM



Component	SMC Part number	Description	Quantity
Barebone	AS -2115HS-TNR	H13SSH, CSE-HS219-R1K63P-A	1
CPU	PSE-GEN9554P-0804	Genoa 9554P UP 64C/128T 3.1G 256M 360W SP5---PSE-GEN9554P-0804	1
Memory	MEM-DR532L-CL02-ER56	32GB DDR5 5600 ECC REG---MEM-DR532L-CL02-ER56	12
Data drives	HDS-MUN-MTFDKCC15T3TFR1B	7450 PRO 15.3TB NVMe PCIe4 3D TLC 15mm,1DWPD	22
Metadata drives	HDS-MUN-MTFDKCC7T6TFR1BC	7450 PRO 7.6TB NVMe PCIe 4.0 3D TLC 15mm,1DWPD	2
Boot drives - M.2	HDS-MMN-MTFDKBA480TFR1BC	Micron 7450 PRO 480GB NVMe PCIe 4.0 M.2 22x80mm 3D TLC, 1DWPD	2
AIOM	AOC-CX7600078-MB0	2-ports NVIDIA 900-9X760-0078-MB0 OCP3.0 200G IB & 2p VPI---AOC-CX7600078-MB0	1
AOC Network	AOC-CX7AH0078-DTZ	900-9X7AH-0078-DTZ NDR 200G 2-port,QSPF112, VPI, No Crypto	1
Accessory	MCP-240-21908-0N	Riser Bracket(4-slot) for RSC-H2-6888G5L in CSE-HS219/HS829,	1
Accessory	CBL-KIT-2115HS-TNR-24N	24 NVME cable kit for AS -2115HS-TNR,RoHS	1
Accessory	RSC-H-6G5L	PCIe	1
Assembly	MC0037	Assembly	1
EWCSC	EWCSC	0% 3 YRS LABOR, 3 YRS PARTS, 1 YR CRS UNDER LIMITED WRNTY	1

NVIDIA networking BOM



Component	SMC Part Number	Description
Switch	SSE-MSN5600-CS2R	Nvidia Spectrum-4 800G 2U Cumulus, 64 OSFP+1 SFP28,C2P
Switch side transceiver	TRX-MMA4Z00-NS	NV 2-p TRX, 800G, 2*NDR,OSFP,2*MPO12 APC, 850nm MMF,50M,Fin
Server side transceiver	TRX-MMA1Z00-NS400	NVDA 1-p TCVR,400G,QSFP112,MPO, 850nm MMF,SR4,30m,flat top
Cable	CBL-SRK-MFP7E20-N015	NVIDIA PFC, MMF, MPO12 APC to 2xMPO12 APC, 15m,RoHS
SuperNIC	GPU-NVDPU-B3140H-C	[DPU]Nvidia BlueField-3 B3140H E-series HHHL DPU, 400GbE---GPU-NVDPU-B3140H-C
DPUs	GPU-NVDPU-BA3220	[DPU] Nvidia BF3 DPU 200G dual-port PCIe Gen5 Crypto Disabled---GPU-NVDPU-BA3220
	GPU-NVDPU-BA3220-C	[DPU] BlueField-3 BF3220 DPU 200GbE dual port Crypto Enable---GPU-NVDPU-BA3220-C
	GPU-NVDPU-BA3240-C	[DPU]NVIDIA BlueField-3 B3240 P-Series Dual-slot FHHL DPU---GPU-NVDPU-BA3240-C

About Cloudian

Cloudian is a Silicon Valley-based file and object storage software company specializing in S3-compatible object storage systems. With technology roots in the large-scale enterprise message space, Cloudian introduced its object-based platform, HyperStore®, in 2011.

About Supermicro

Supermicro is a global technology leader committed to delivering first-to-market innovation for Enterprise, Cloud, AI, Metaverse, and 5G Telco/Edge IT Infrastructure. Supermicro is a Rack-Scale Total IT Solutions provider that designs and builds an environmentally friendly and energy-saving portfolio of servers, storage systems, switches, software, along with global support services. As a leader in energy-efficient computing, Supermicro's goal is to promote the adoption and deployment of technologies that can reduce costs as well as the impact on the environment. Resource Saving Architecture continues our tradition of green computing innovation and provides TCO savings for our customers. https://www.supermicro.com/en/about