WEKA Qualified Program Nitro Reference Design for NVIDIA GB200/GB300 NVL72 Systems

This white paper presents a validated WEKA reference design for NVIDIA Cloud Partner high-performance storage deployments using NeuralMesh, WEKA-qualified storage servers, and NVIDIA GB200/GB300 NVL72 systems. The design is built to meet NCP HPS requirements by combining a scale-out parallel file system, and low-latency networking to deliver the throughput, latency, and scalability needed for large AI training and inference environments.

It emphasizes operational simplicity by running metadata and data services on each storage server, supporting a single namespace that scales from 8 servers to hundreds, while also providing enterprise features such as multitenancy, encryption, snapshots, tiering, backup, and QoS.

The paper also includes qualified hardware specifications, networking and cabling guidance, and storage sizing recommendations ranging from 1,152 GPUs to 41,472 GPUs, positioning WEKA as a scalable, production-ready storage foundation for large NVIDIA AI cloud deployments.

Table of Contents:

1. Executive Summary

NeuralMesh™ software, combined with class-leading storage hardware, provides a ready-to-use, purpose-built environment specifically designed for artificial intelligence (AI) applications.

WEKA empowers hundreds of the world's leading enterprises and premier research organizations to overcome complex data challenges, enabling them to achieve discoveries, insights, and outcomes more rapidly. This includes 12 of the Fortune 50 companies that rely on WEKA to enhance their data-driven initiatives.

This document outlines a validated reference design for NVIDIA® Cloud Partners (NCPs). The NCP Solution is tested with NeuralMesh, integrating the WEKA Storage Servers and WEKA Management Station (WMS) software, and NVIDIA GB200 and GB300 NVL72 systems in alignment with the NVIDIA NCP High-Performance Storage (HPS) tier requirements.

With the completion of this NCP HPS Reference Design, WEKA and NVIDIA extend their deeper technical integration collaboration that includes development and certification with NVIDIA Magnum IO™ GPUDirect Storage® (GDS), NVIDIA DGX BasePODTM, NVIDIA DGX SuperPODTM, NVIDIA OVX™ systems, NVIDIA networking solutions including NVIDIA ConnectX® NICs, NVIDIA Quantum InfiniBand and Spectrum™ Ethernet Platforms, and NVIDIA BlueField®-3 DPU/SuperNICs.

2. NeuralMesh for NVIDIA Cloud Partners

NVIDIA GB200 NLV72 and GB300 NVL72 systems are engineered to deliver unparalleled performance for AI and machine learning (ML) workloads. These systems enable organizations to optimize their infrastructure for specific AI and data-intensive applications. The integration of NeuralMesh with NVIDIA GB200 and GB300 systems elevates performance to new heights, providing NVIDIA Cloud Partners with an unmatched data processing and storage experience for AI and ML workloads.

Starting with a base configuration of eight servers, NeuralMesh scales linearly and seamlessly to hundreds of servers, providing capacity, throughput, and IOPS expansion as needed. NeuralMesh delivers best-in-class read/write and mixed-workload performance, supporting both small and large files simultaneously, with mixed random and sequential I/O patterns. It achieves application-level 4kB I/O at sub-200 microsecond latency and tens of millions of IOPS, ensuring data is readily accessible for demanding AI workloads.

Designed to complement the computational power of NVIDIA GB200 and GB300 systems, NeuralMesh provides seamless, high-performance data access and storage with enterprise-grade

features. The NCPs benefit from NeuralMesh's rich feature set, including local snapshots, automated tiering, dynamic cluster rebalancing, private cloud multitenancy, backup, encryption, authentication, key management, user groups, quotas, and more. By leveraging NeuralMesh on NVIDIA GB200 and GB300 systems, enterprises can optimize their AI pipelines and accelerate time-to-insight across a wide range of use cases.

This section describes the advanced features of NeuralMesh for NCPs.

2.1 NeuralMesh Architecture

NeuralMesh's containerized microservices architecture is purpose-built to deliver high-performance file services leveraging NVMe flash in a single namespace for easy access and management.

NeuralMesh's unique architecture is radically different from legacy storage systems, appliances, and hypervisor-based software-defined storage solutions because it not only overcomes traditional storage scaling and file sharing limitations but also allows parallel file access via POSIX, particularly using the Data Plane Development Kit (DPDK). The DPDK is a set of libraries and drivers for fast packet processing. Using DPDK, NeuralMesh can achieve high-speed data processing and low latency, making it ideal for performance-critical applications. This allows for efficient parallel file access, significantly enhancing the system's overall performance and scalability. It provides a rich enterprise feature set, including local snapshots and remote snapshots to the cloud, clones, automated tiering, cloud-bursting, dynamic cluster rebalancing, private cloud multitenancy, backup, encryption, authentication, key management, user groups, quotas with advisory, soft and hard parameters, and much more.

NeuralMesh's patented data layout and virtual metadata servers distribute and parallelize all metadata and data across the cluster for incredibly low latency and high performance no matter the file size or number.

At the core of NeuralMesh is WekaFS™, a distributed, parallel file system that eliminates the traditional block-volume layer managing underlying storage resources. WekaFS is an advanced file system designed to overcome the limitations of traditional storage solutions by offering superior performance, scalability, and data integrity. It features adaptive caching, which optimizes the use of local Linux data and metadata caches to significantly reduce latency and improve performance. WekaFS ensures full coherence and consistency across the shared storage cluster, automatically managing caching to eliminate the need for specific configurations or administrative interventions. This makes it an ideal choice for performance-critical applications, ensuring data integrity and ease of management.

The WEKA core components, including the WekaFS unified namespace and other functions such as virtual metadata servers (MDSs), execute in user space in a Linux container (LXC), effectively

eliminating time-sharing and other kernel-specific dependencies. The exception is the WEKA Virtual File System (VFS) kernel driver, which provides the POSIX filesystem interface to applications. Using the kernel driver provides significantly higher performance than what can be achieved using a FUSE user-space driver, and it allows applications that require full POSIX compatibility to run on a shared storage system.

2.2 NeuralMesh Performance-Optimized Networking (DPDK)

To achieve optimum performance, NeuralMesh does not use standard kernel-based TCP/IP services but instead employs a proprietary networking stack based on DPDK maps the network device in the user space, allowing the network device to be used without context switches or copying data between kernels. Bypassing the kernel stack eliminates the consumption of kernel resources for networking operations and efficiently scales across multiple hosts. It applies to both backend and client hosts and enables the WEKA system to fully use 400 GbE links.

Using DPDK provides operations with high throughput and extremely low latency. Low latency is achieved by bypassing the kernel and sending and receiving packets directly from the NIC. High throughput is achieved because multiple cores in the same host can work in parallel, eliminating common bottlenecks.

2.3 NeuralMesh Multitenancy, Security, and QoS

A collection of capabilities provides multitenancy for NeuralMesh. Tenant isolation is both physical and logical, with each tenant being a self-contained NeuralMesh instance. Tenant creation is governed by the WEKA Kubernetes (K8s) Operator (“Operator”) through composable sets of resources—these composable clusters are deployed adjacently to each other in collections of physical servers. The fundamental resource allocation method for tenants is dedicated components and physical isolation, allowing for naturally gated performance and strong security boundaries without extensive management and QoS overhead. The tenants do not share flash drives, drive memory, processor cores, or main memory.

Each tenant uses unique encryption keys and authentication credentials and is unaware of each other despite being adjacent. Encryption is set at the file system level upon creation of the file system, so file systems that are deemed critical can be encrypted while others are not. When files are encrypted, it means that even if cold blocks underlying the file are sent to an object store tier, the data will remain encrypted. NeuralMesh’s encryption solution protects against physical media theft, low-level firmware hacking on the SSD, and packet eavesdropping on the network. File data is encrypted with the FIPS 140-3 Level 1 compliant encryption key XTS-AES using a 512-bit key length. Clients do encryption and decryption, with the native data structures in WEKA containing wrapped key material. The KMS configuration for the cluster enables a key for each file system, all file data, and sensitive file metadata such as the file name. The only thing not encrypted is the size of the file.

For service providers, this is a departure from the traditional storage approach to multitenancy, leveraging hierarchical relationships and weaker process isolation with shared memory and contextual permissions by the tenant. While the Organizations feature of WEKA NeuralMesh also supports this model for trusted tenants and logical delegation hierarchies, dedicating resources allows for strong performance guarantees and security domain separation between tenants throughout the control and data paths, a necessary capability for service providers with untrusted tenants.

Logical isolation includes the network, as it is the only shared part. It can be changed per tenant using virtual functions to control bandwidth consumption by cluster and QoS settings by client. Using the Operator, a cluster can be expanded incrementally to add drives, cores, or a combination of both. The multitenancy capability scales up to 150,000 total tenants.

NeuralMesh deployments leverage templates with composable resource sets for achieving desired capacity and performance in a tenant. Every drive and core will produce a certain amount of performance, and the arithmetic is handled by the Operator. The mechanics are that the Operator allocates cores and memory in physical servers to K8s containers, which contain LXC containers used by NeuralMesh. Software inside the LXC containers forms a cluster and leverages all the previously mentioned performance mechanisms like DPDK. Next, the Operator adds drives to the tenant NeuralMesh creation is fast, with most tenant clusters, wall clock time to first I/O will be measured in single-digit minutes. Administrators can deploy both small and large sets of capacity with guaranteed performance as shown in Figure 1.

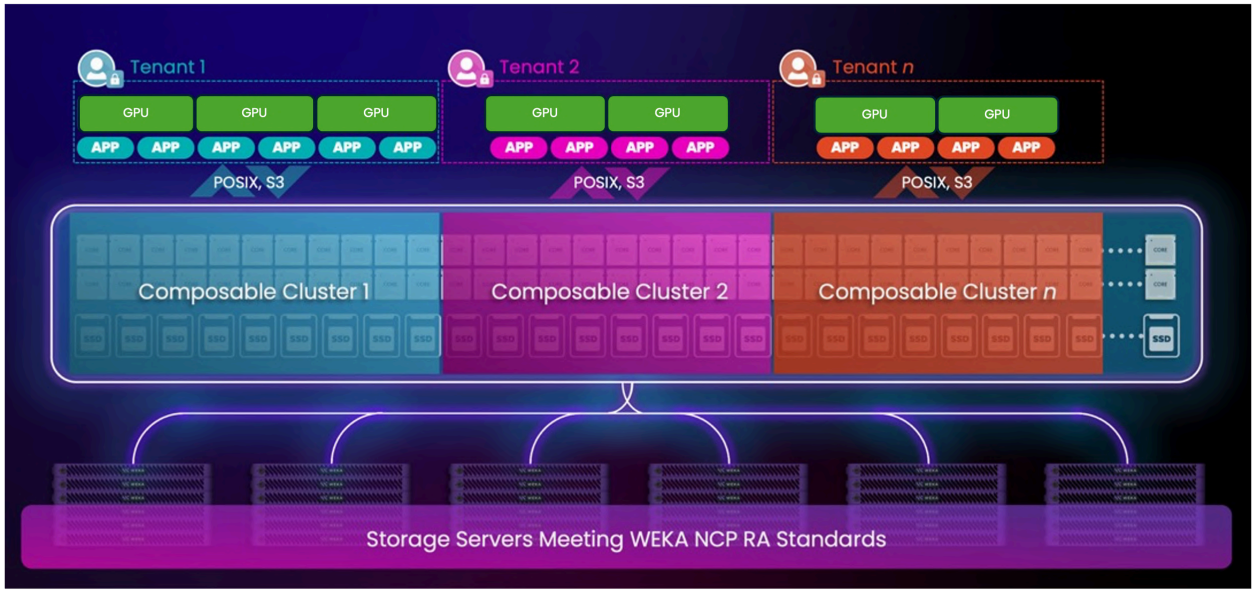

Figure 1. WEKA multitenant cluster architecture design

3. WEKA Storage Server Qualified Configuration

The WEKApod appliance or any of the WEKA qualified platform from other vendors meeting the specification listed in Table 1 can be used as WEKA storage:

Hardware Specifications	
Processor	AMD EPYC 9534 64C 2.45GHz Processor
Memory	Min. 768 GB (1 DIMM per Memory Channel). DIMM is DDR5 RDIMM 4800MTs Dual Rank
Storage	14 or more E3.s Gen5 NVMe TLC SSD. A minimum of 40x PCIe 5.0 lanes to the drives
Network	2 x NVIDIA ConnectX-7 1P 400 Gb/s OSFP PCIe 5.0 (requires 2x PCIe 5 x16 I/O Slots)

Table 1. WEKA storage server specifications used during NCP validation

The WEKA storage server integrates WEKA's fully distributed parallel file server architecture and includes NeuralMesh software, as described in Section 1, providing a ready-to-use, purpose-built environment for AI applications. Each WEKA storage server runs both metadata and data services, simplifying NCP deployments. NeuralMesh interfaces with GB200 and GB300 systems using DPDK, a highly efficient and low-latency packet processing technology, over ConnectX®-7 NICs. Each WEKA storage server provides two ConnectX-7 NICs to connect to the storage fabric. During the NCP Reference design validation, WEKA used and tested the WEKA Storage Server™ with specifications listed in Table 1.

3.1 NeuralMesh Management GUI

WEKA provides three quick and easy ways to manage the WEKA file system, either through a Graphical User Interface (GUI), or a Command Line Interface (CLI), or a representational State Transfer API (REST). Reporting, visualization, and overall system management functions are accessible using the REST API, CLI, or the intuitive GUI-driven management console (see Figure 2 and Figure 3). Point-and-click simplicity allows users to rapidly provision new storage, create and expand file systems within a single namespace, establish tiering policy, data protection, encryption, authentication, permissions, NFS, SMB and S3 configuration, read-only or read-write snapshots, snapshot-to-objects, and quality of service policies, as well as monitor overall system health.

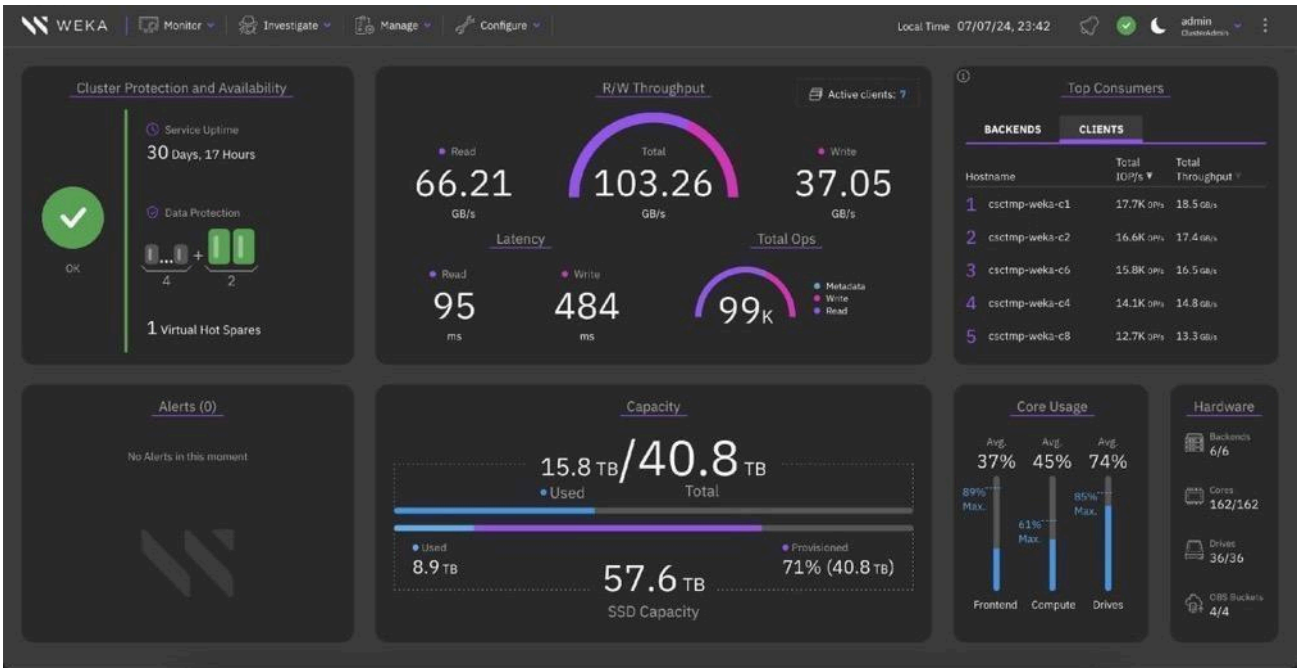

Figure 2. WEKA management software user interface.

Detailed event logging provides users with the ability to view system events and status over time or drill down into event details with point-in-time precision via the time-series graphing function.

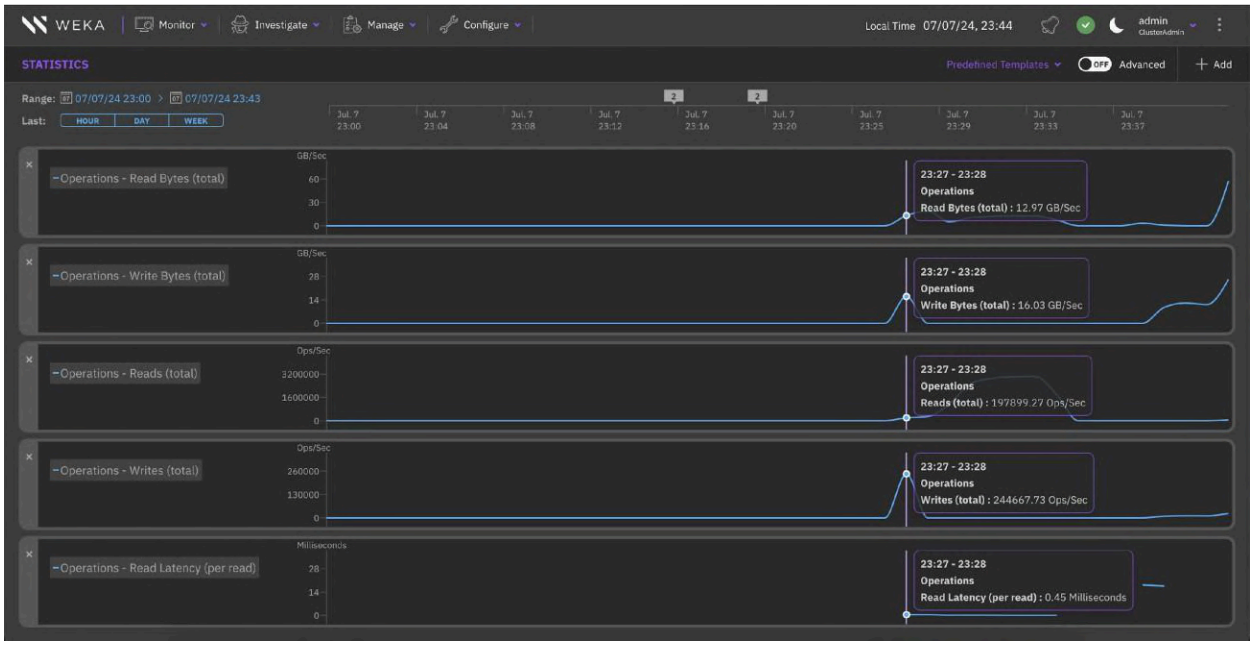

Figure 3. Time-series charts for event monitoring.

4. WEKA Reference Design for NVIDIA NCP Deployments

WEKA qualified Storage Servers deliver all the capabilities of NeuralMesh software in a predictable scale-out reference design that's easy to deploy. NeuralMesh offers an exceptional, highly performant data management foundation for NCP deployments. NeuralMesh also provides efficient write performance for AI model checkpointing, delivering superior training efficiency, scalability, enterprise-grade resiliency, and support for real-time applications. WEKA Storage Servers provide both metadata and data services, streamlining the design, deployment, and management of a cloud environment, offering predictable performance, capacity, and scalability. WEKA proposes the following recommended architectures for NCP deployments.

4.1 Recommended Storage Sizing

The storage performance target for training or inference can vary depending on the type of model and dataset. The guidelines in Table 2 provide standard throughput for the various GPU system sizes and HPS sizing. The final HPS requirements for throughput and capacity will be specified for each NCP opportunity.

Description	Number of GPUs						
	1,152	2,304	4,608	8,064	16,128	29,952	41,472
Read throughput (GB/s)	180	360	720	1,260	2,520	4,680	6,480
Write throughput (GB/s)	90	180	360	630	1,260	2,340	3,240
Storage Configuration							
Number of Storage Servers	8	8	11	18	36	67	93
Number of namespaces	1	1	1	1	1	1	1
Number of 400G storage ports	16	16	22	36	72	134	186
Number of in-band management connections	8	8	11	18	36	67	93
Number of out-of-band management connections	8	8	11	18	36	67	93
Number of rack units	9	9	12	19	37	68	94
Power (KW)	6.4	6.4	8.8	14.4	28.8	53.6	74.4
Cooling (BTU/hr)	22,000	22,000	30,250	49,500	99,000	184,250	255,750

Table 2. Guidance for standard HPS aggregate storage performance

4.2 WEKA Storage Networking

WEKA Storage Servers integrate seamlessly with NVIDIA Spectrum Ethernet SN5600/5610 leaf switches using 16 x 400 Gb connections per 8-node cluster. The platform fully supports DPDK to enable direct data transfers from storage to GPU memory, minimizing CPU overhead and maximizing bandwidth efficiency.

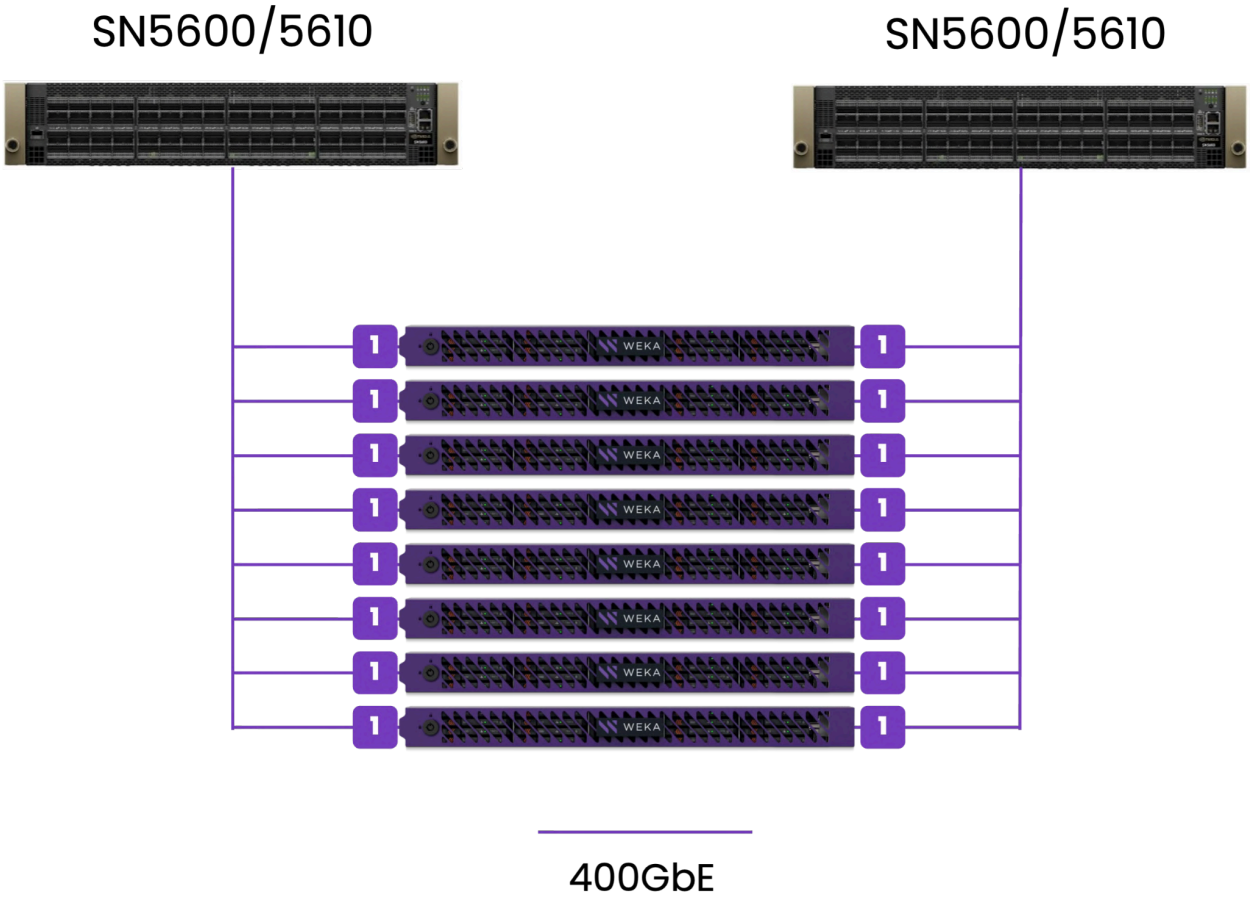

Figure 4. WEKA Storage Servers and SN5600/5610 Leaf switch connectivity overview

4.3 WEKA Storage Server Cabling with SN5600/5610 Switches

An NVIDIA MMS4X00-NS twin-port 2x400Gb/s OSFP transceiver is utilized with NVIDIA Spectrum Ethernet SN5600/5610 leaf switches to connect WEKA Storage Servers. Each WEKA storage server includes two NVIDIA ConnectX-7 NICs and employs NVIDIA MMS4X00-NS400 transceivers to establish connections to a pair of switches in an active-active configuration.

WEKA and NVIDIA have jointly validated the cables listed in Table 3 for connecting WEKA Storage Servers to NVIDIA Spectrum Ethernet SN5600/5610 switches. Using splitter cables optimizes switch port utilization.

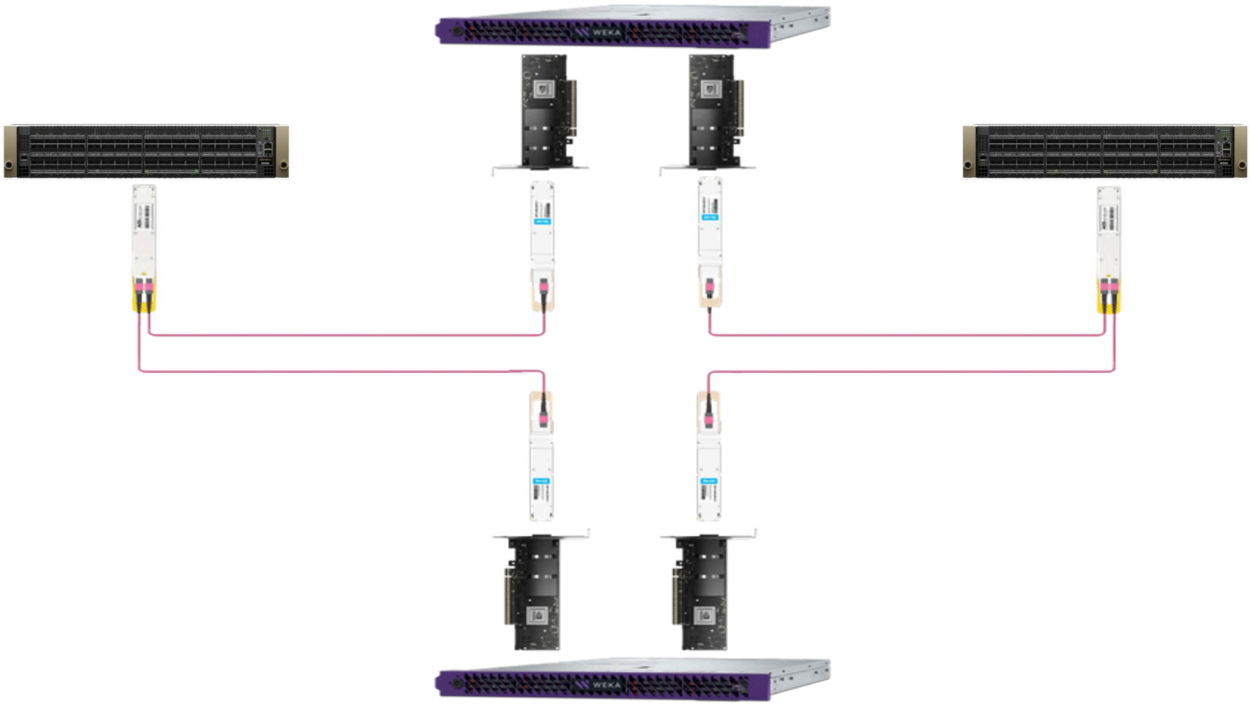

Figure 5. Transceivers and cabling connecting WEKA Storage Servers and NVIDIA Spectrum Ethernet SN5600/5610 leaf switches

Processor	Processor	Description
Direct Attach Copper	MCP7Y00-Nxxx	NVIDIA 800 Gb/s Twin-port OSFP to 2x400 Gb/s OSFP DAC Splitter Cable xxx indicates length in meters: 001, 01A, 002, 02A, 003
Active Copper Cable	MCA7J60-Nxxx	NVIDIA 800 Gb/s Twin-port OSFP to 2x400 Gb/s OSFP ACC Splitter xxx indicates the length in meters: 004, 005

Table 3. DAC and ACC splitter cables

The Figure 6 illustrates the network connectivity of WEKA Storage Servers to a pair of NVIDIA Spectrum Ethernet SN5600/5610 leaf switches, establishing a high-throughput, non-blocking network architecture.

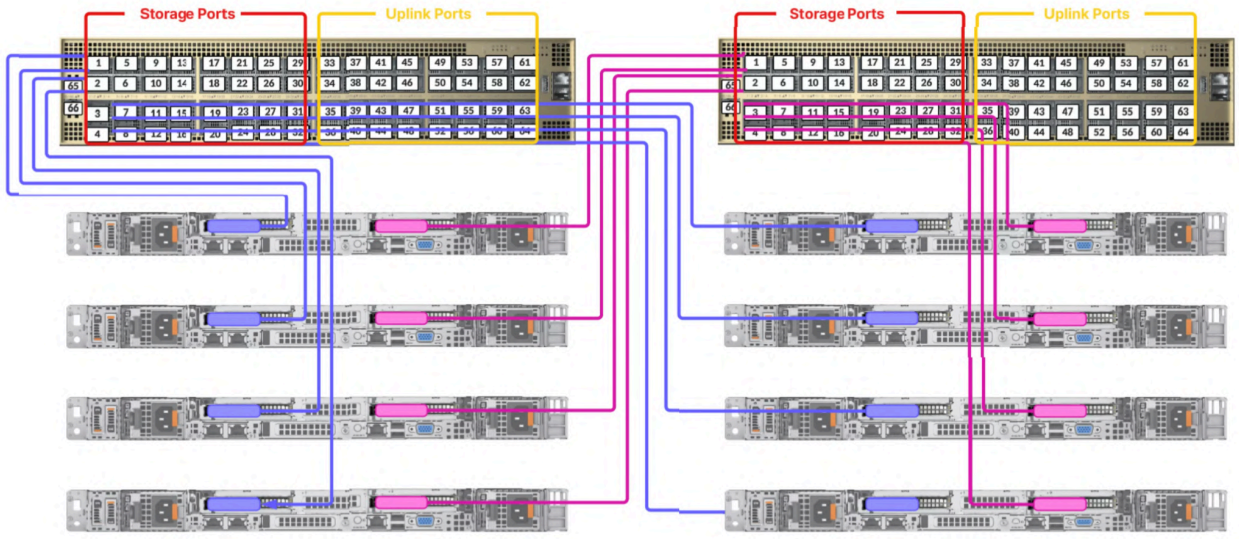

Figure 6. WEKA storage server port mapping to NVIDIA Spectrum Ethernet SN5600/5610 leaf pair

4.4 NCP Deployments with 1,152 GPUs

Figure 7 illustrates WEKA's reference design for NCP deployments with 1,152 GPUs, 8 x WEKA Storage Servers. Every WEKA storage server connects to the NVIDIA Spectrum Ethernet SN5600/5610 leaf switch network with two 400 GbE links using the appropriate cable type as listed in Table 4.

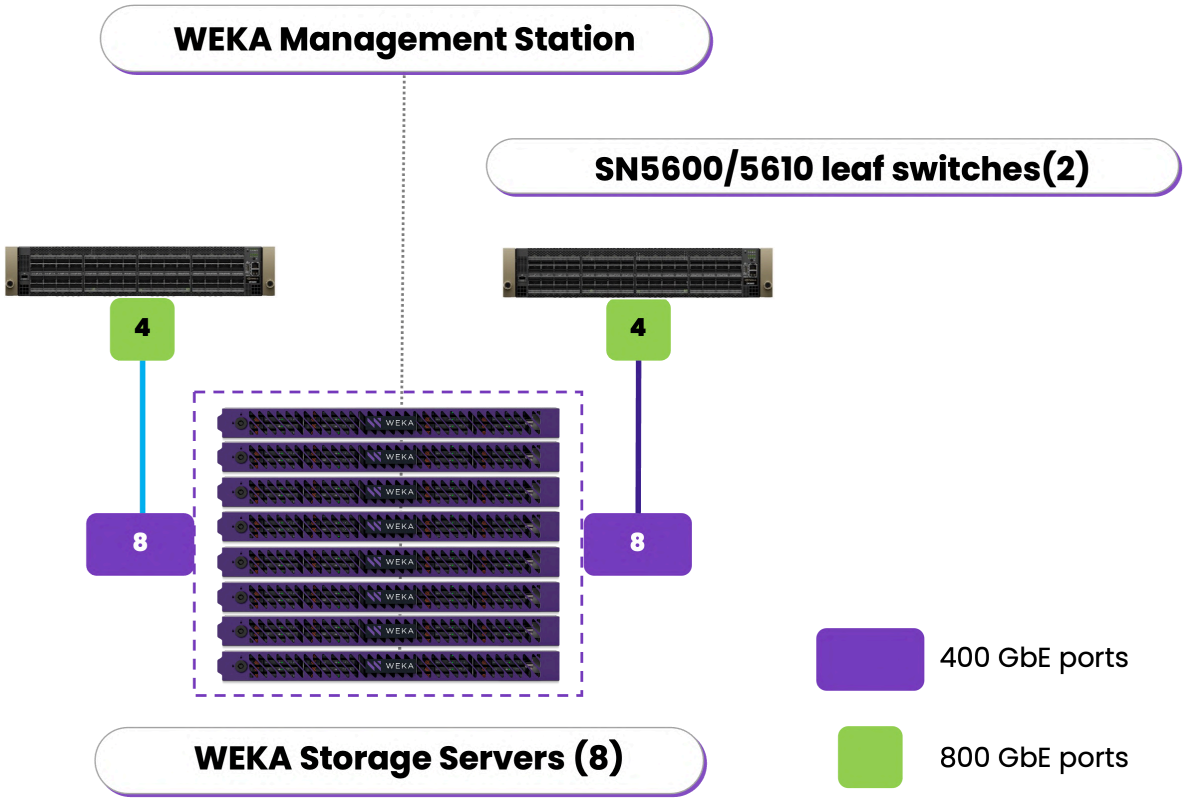

Figure 7. WEKA Storage Servers Reference Design for NCP deployments with 1,152 GPUs.

Cable counts for NCP deployments with 1,152 GPUs	
Processor	**Count**
400 GbE links (Storage)	16
OSFP ports (Switch) / Splitter cables (MCP700-Nxxx)	8
1 GbE links (Management)	8

Table 4. Cable counts for NCP deployments with 1,152 GPUs

4.5 NCP Deployments with 16,128 GPUs

Figure 8 illustrates WEKA's reference design for NCP deployments with 16,128 GPUs, 36 x WEKA Storage Servers. Each WEKA storage server connects to the storage network with two 400GbE links using the appropriate cable type as listed in Table 5.

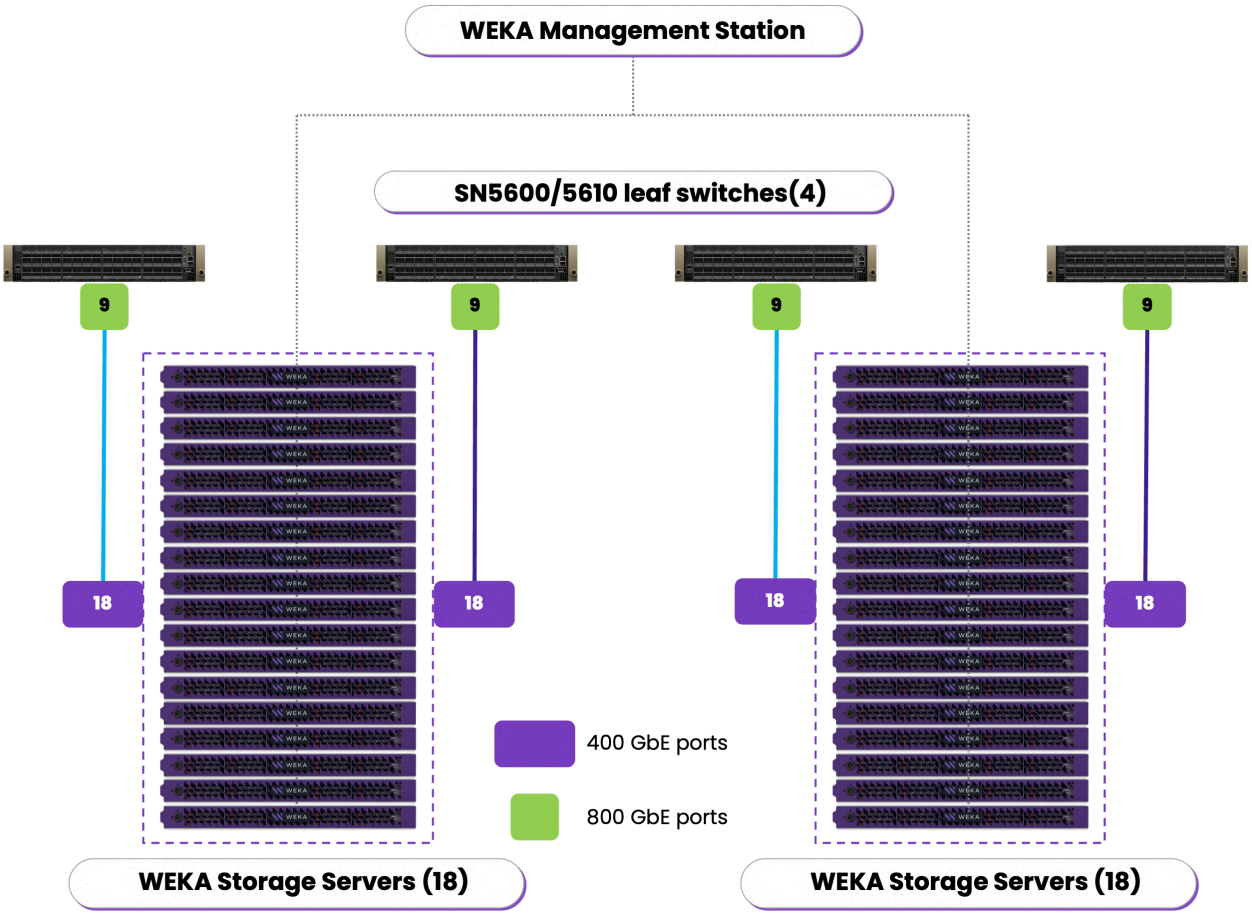

Figure 8. WEKA Storage Servers Reference Design for NCP deployments with 16,128 GPUs.

Cable counts for NCP deployments with 1,152 GPUs	
Processor	**Count**
400 GbE links (Storage)	72
OSFP ports (Switch) / Splitter cables (MCP700-Nxxx)	36
1 GbE links (Management)	36

Table 5. Cable counts for NCP deployments with 16,128 GPUs

4.6 NCP Deployments with 41,472 GPUs

Figure 9 illustrates WEKA's reference design for NCP deployments with 41,472 GPUs, 93 x WEKA Storage Servers. Each WEKA storage server connects to the storage network with two 400 GbE links using the appropriate cable type as listed in Table 6.

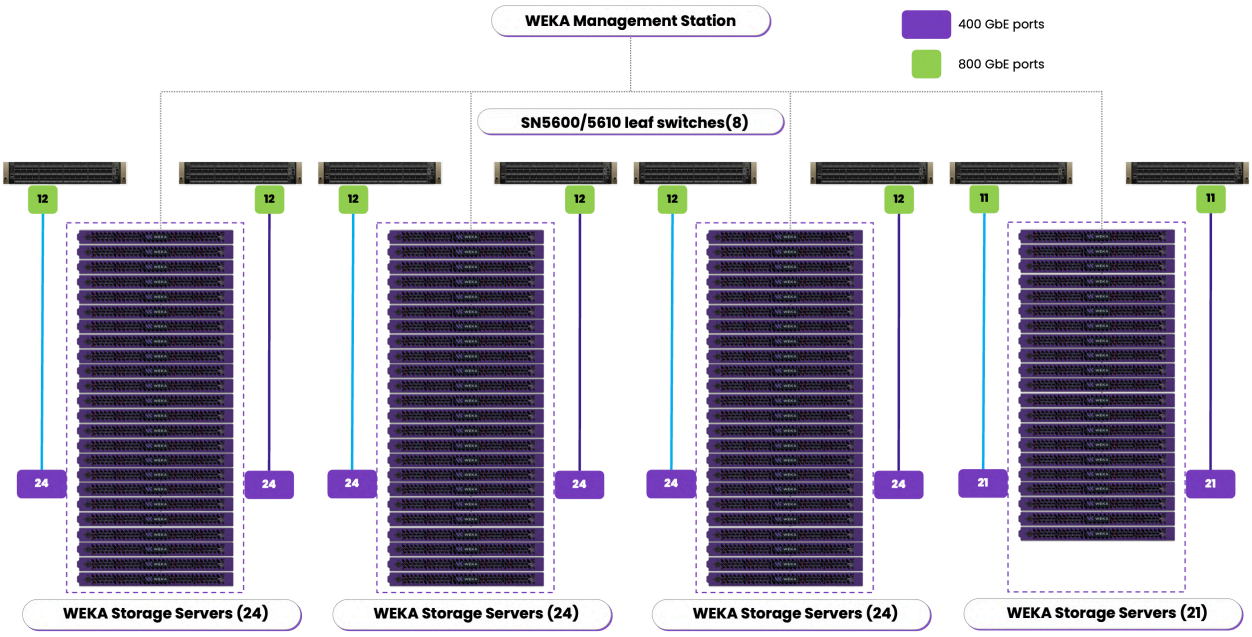

Figure 9. WEKA Storage Servers Reference Design for NCP deployments with 41,472 GPUs

Cable counts for NCP deployments with 1,152 GPUs	
Processor	**Count**
400 GbE links (Storage)	186
OSFP ports (Switch) / Splitter cables (MCP700-Nxxx)	94
1 GbE links (Management)	94

Table 6. Cable counts for NCP deployments with 41,472 GPUs

5. Summary

NeuralMesh has pioneered an architecture that not only delivers outstanding performance and scalability but also ensures operational efficiency and ease of use. This document presents a validated reference design for NeuralMesh, specifically adhering to the NCP RA specifications for HPS. With flexible configurations supporting over 41K GPUs in a single cluster, NCPs can confidently pair NeuralMesh with large-scale AI infrastructure deployments. The jointly validated NeuralMesh reference design, based on the NCP RA and featuring GB200 NVL72 or GB300 NVL72 systems and NVIDIA Spectrum Ethernet networking, meets and exceeds the rigorous demands of AI, HPC, and other data-intensive applications.

6. Appendix A: Data Services & Unified Security Model

NeuralMesh supports multi-protocol and data-sharing capabilities across various protocols, allowing diverse application types and users to share a single pool of data. Unlike other parallel file systems, WEKA does not require a gateway server infrastructure to offer this capability, as shown in Figure 10.

The currently supported protocols include:

- Full POSIX for local file system support
- NFS for Linux
- SMB for Windows
- S3 for Object access

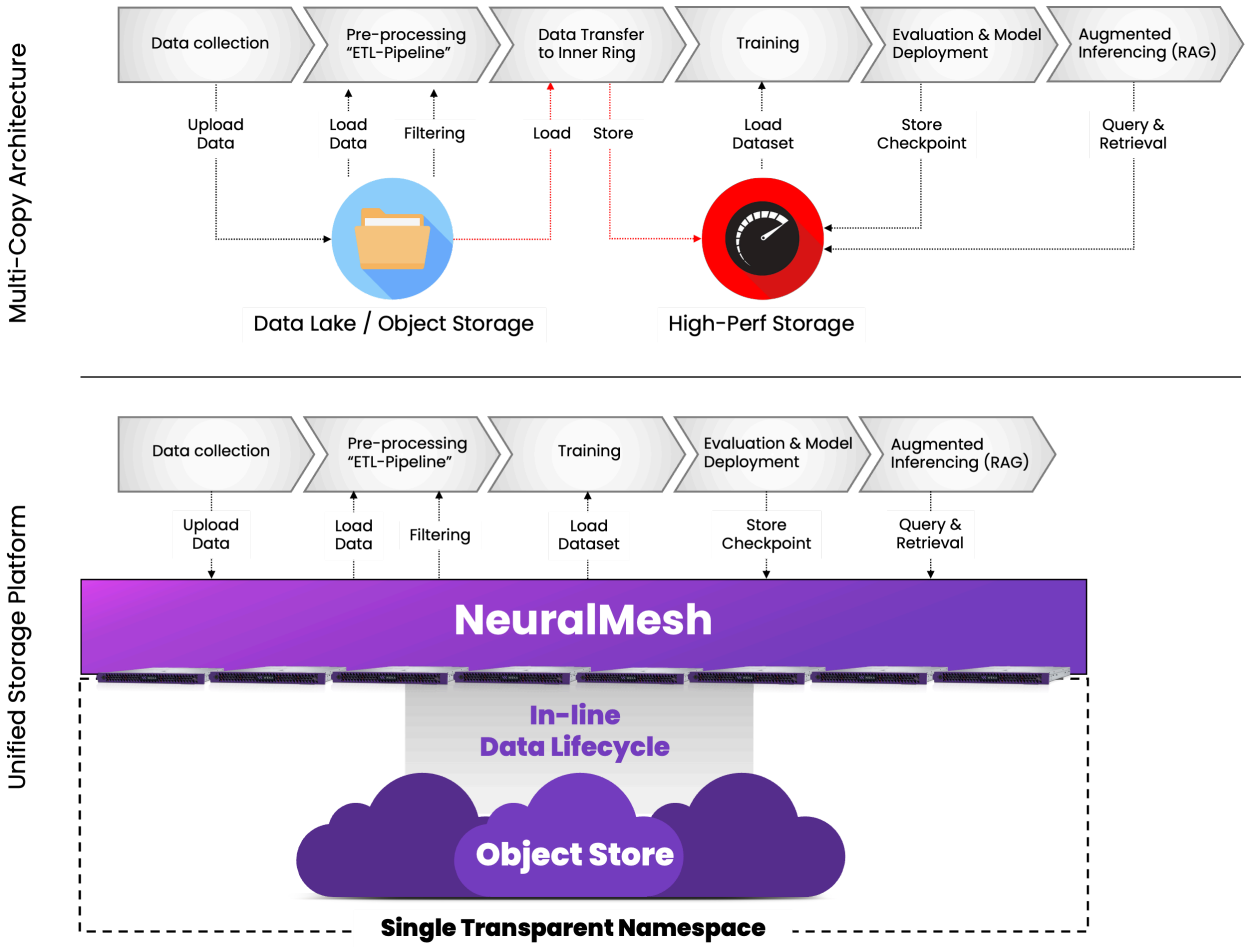

Figure 10. WEKA NeuralMesh, providing data access to every stage of the AI pipeline

By supporting POSIX and file/object protocols, NeuralMesh allows organizations to consolidate HPS (training data) and data lake (data collection and preprocessing) storage needs onto a single platform. This consolidation eliminates the need for data transfers between separate storage systems, thereby simplifying AI training, inference, and model serving processes. This integration enhances speed and reliability, streamlining the entire AI data lifecycle. Consequently, organizations can focus on innovation rather than dealing with intricate multi-copy storage architectures and burdensome data transfer operations.

Furthermore, by unifying various storage systems into a single, cohesive storage system, WEKA empowers NVIDIA Cloud Partners to offer a highly differentiated solution that accelerates the entire AI pipeline. This capability enables NCPs to develop and deliver services to their customers or users more quickly and with reduced operational overhead, in contrast to traditional multi-copy architectures.

7. Appendix B: WEKA configuration for 1 GB/s per GPU of performance

The High Performance Storage network in NVIDIA GB200 and GB300 based systems can easily handle up to 1 GB/s per GPU. The Table 7 details the sizing guidance for WEKA storage cluster to meet the NCP deployments with 1.0 GB/s of read per GPU requirement for respective scalable unit configurations.

Description	Number of GPUs						
	1,152	2,304	4,608	8,064	16,128	29,952	41,472
Read throughput (GB/s)	1152	2304	4608	8064	16128	29952	41472
Write throughput (GB/s)	576	1152	2304	4032	8064	14976	20736
Storage Configuration							
Number of Storage Servers	17	33	66	116	231	428	593
Number of namespaces	1	1	1	1	1	1	1
Number of 400G storage ports	34	66	132	232	462	856	1186
Number of in-band management connections	17	33	66	116	231	428	593
Number of out-of-band management connections	17	33	66	116	231	428	593
Number of rack units	18	34	67	117	232	429	594
Power (KW)	13.6	26.4	52.8	92.8	184.8	342.4	474.4
Cooling (BTU/hr)	46,750	90,750	181,500	319,000	635,250	1,177,000	1,630,750

Table 7 . WEKA recommended storage sizing for NCP deployments with 1.0 GB/s of read per GPU