

HPE Solutions with WEKA for NVIDIA Cloud Partners on HPE ProLiant DL325 Gen11



What are NVIDIA Cloud Partners and why they matter

NCPs are a curated group of cloud providers and infrastructure partners that collaborate with NVIDIA to deliver optimized solutions for AI, ML, and data-intensive workloads.

- **Built for performance:** These partners offer cloud solutions designed to maximize the performance of NVIDIA GPUs and technologies, for efficient deployment and scaling of AI workloads.
- **Standards-driven reliability:** NCPs meet rigorous performance and compliance standards, in areas such as GPU acceleration, HPS, and networking.
- **Advanced AI Infrastructure:** Organizations can leverage NVIDIA's advanced AI platforms such as HGX systems, GPUDirect Storage, and Magnum IO.

Purpose-built for hybrid and public cloud AI environments, NeuralMesh™ by WEKA on the HPE ProLiant DL325 Gen11 delivers exceptional performance that meets NVIDIA® Cloud Partner (NCP) high-performance storage (HPS) specifications.

NCP-qualified storage with HPE + WEKA High-performance AI infrastructure

HPE and WEKA have partnered to deliver industry-leading server innovation with a flash-native data platform to provide scalable, high-throughput, low-latency AI storage. Designed to support large-scale and data-intensive workloads, this solution offers the scale, efficiency, and speed modern AI applications require.

NeuralMesh by WEKA on the HPE ProLiant DL325 Gen11 is a WEKA-qualified platform designed to meet NVIDIA NCP reference architecture compliance. This economical, 1U 1P system balances compute, memory, and network bandwidth, for exceptional performance at a lower cost.

Product overview

- WEKA-certified reference architecture built on HPE ProLiant DL325 Gen11 with PCIe Gen5 performance, NVIDIA CX7-NDR networking, and EDSFF drive technology
- Powered by WEKA's NeuralMesh
- NVMe and NVMe-oF technology (remote direct memory access [RDMA], and TCP)
- Magnum IO GPUDirect Storage (GDS)
- POSIX, NFS, S3, and SMB protocol support

Platform performance

(8-node cluster)

Throughput

- 720 GB/s sustained read bandwidth
- 256 GB/s sustained write bandwidth

IOPS

- 18.3 million 4 KB random read IOPS
- 4.3 million 4 KB random write IOPS

Latency

- 162µs 4 KB read latency
- 116µs 4 KB write latency

Simplicity

- REST API for orchestration
- Nondisruptive upgrades (NDU) and capacity expansions
- Native cloud burst to all major hyperscalers (AWS, Azure, GCP™, Oracle® Cloud Infrastructure)
- NVIDIA Base Command Manager integration

Scalability

- Scale to 100s of servers
- Linear TB, BW, IOPS at-scale
- Trillions of inodes per deployment
- Billions of inodes per directory tree

Sustainability

- 10x to 50x better stack efficiency¹
- 4x to 7x lower data center footprint²
- 260 tons of CO₂e saved per PB annually versus traditional storage³

HPE ProLiant DL325 Gen11: Optimized for WEKA

Backed by HPE's legacy of innovation and performance leadership, the HPE ProLiant DL325 Gen11 server is a 1U, single-socket platform powered by 4th Gen AMD EPYC™ processors. With expanded memory bandwidth, high-speed PCIe Gen5 I/O, and EDSFF storage support, it delivers outstanding throughput, IOPS, and low latency.

Jointly tested and validated by HPE and WEKA, this system accelerates data workflows, reduces epoch times, and enables faster inferencing—unlocking the full potential of WEKA's NeuralMesh.

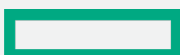
NeuralMesh™ by WEKA: Purpose-built to accelerate AI at scale

NeuralMesh is a modern, software-based data platform that transforms enterprise data stacks by eliminating traditional storage bottlenecks to GPU performance. The solution supports parallel file access across POSIX, NFS, SMB, S3, and NVIDIA Magnum IO GPUDirect Storage (GDS), combining SAN-like low-latency with object store-level scalability for ultra-fast, consistent performance.

The platform supports a rich enterprise feature set, including snapshots, automated tiering, dynamic cluster rebalancing, multitenancy, and comprehensive security features such as encryption and key management.

Cloud and hardware-agnostic, WEKA provides seamless data portability across on-premises, edge, and hybrid/multicloud environments, enabling organizations to run workloads efficiently wherever needed. Available as a subscription, fully managed service, or a turnkey appliance, WEKA offers flexible deployment options to meet the needs of modern enterprises.

^{1, 2, 3} ["Save 260 tons of CO2e per PB with the WEKA Data Platform,"](#) WEKA, 2023





Feeding GPUs from a single namespace

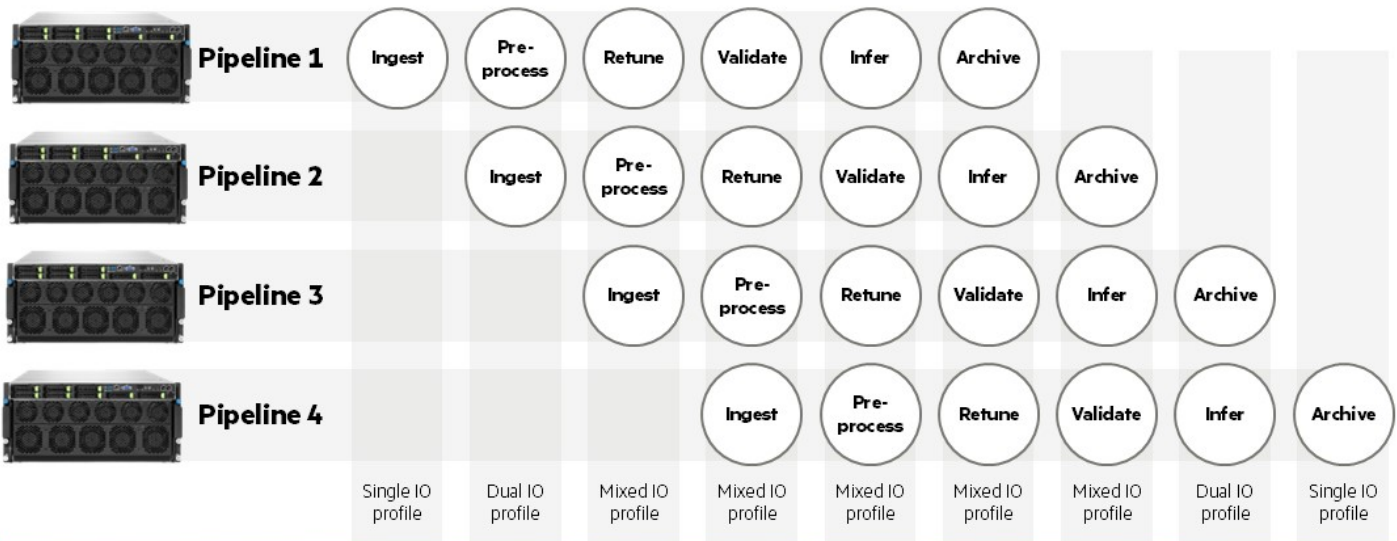
WEKA optimizes typical GPU-starved AI pipelines by storing your entire dataset in a single namespace, removing the need for inefficient multi-hop staging. This zero-copy architecture gives GPUs fast, direct access to training data, automated tiering seamlessly moves datasets between high performance, NVMe-based storage, and low-cost object storage.

Incorporating NeuralMesh into deep learning data pipelines dramatically increases the utilization rates for NVIDIA GPU systems. Eliminating any redundant data movement and transfer times between storage silos significantly boosts the volume of training datasets processed daily.

Delivering multidimensional AI performance

When AI data pipelines are overlapped, a storage system no longer needs to deal with just changing I/O for every step in a pipeline. It now must handle the mixed I/O from different stages of every pipeline and regular checkpoints for save states. Starting parallel training and/or retuning jobs at different times blends I/O patterns to the point where the storage system is dealing with a mixed I/O profile that tends to be random in nature, slowing it down significantly.

WEKA is engineered to handle this I/O blender with ease. It supports small and large files simultaneously, with both mixed random and sequential I/O patterns while delivering application level 4 KB I/O at sub-200µ second latency and tens of millions of IOPS. The starting WEKA configuration delivers up to 720 GB/s of sustained read bandwidth, 186 GB/s of sustained write bandwidth, and up to 18.3 million random 4 KB IOPS.



NeuralMesh™ by WEKA

Zero Copy, Zero Tuning Architecture



Distributed data protection



Intelligent fast rebuild



Instant snapshots & snap clones



Snap & tier to S3 object store



Multiprotocol NFS, SMB, S3, POSIX



Backup/DR & Cloud Burst



AI-rest & in-flight encryption

Figure 1. NeuralMesh by WEKA’s zero-copy, zero-tuning architecture and focus on low latency handles high IOPs or mixed workflows across parallel data pipelines without performance tradeoffs





Fully integrated, GPU-optimized platform

HPE Solutions with WEKA enables the full range of benefits of NeuralMesh by WEKA in a validated, WEKA-certified data platform appliance that uses HPE servers. Starting at a minimum cluster size of 8 nodes, the solution can scale to hundreds of nodes and seamlessly integrates with Run.ai and other orchestration tools. It enables consistent, high-performance data access across distributed environments, helping enterprises keep pace with growing AI demands.

Key features of HPE ProLiant DL325 Gen11

Performance

Advanced data transfer rates and higher network speeds are enabled through the PCIe Gen5 serial expansion bus, with up to 2 x16 PCIe Gen5, two OCP 3.0 slots, two PCIe slots, and improved I/O throughput while also reducing latency. HPE has designed these systems with an optimal balance of CPU clock speed, cores, memory, network performance, and drive performance—all engineered to maintain the highest performance even within tight thermal constraints.

Scale

HPE Solutions with WEKA delivers performance that scales linearly as you add nodes or drives. High-speed storage and large amount of I/O allows you to take full advantage of the AMD infrastructure. The DL325 Gen11 supports up to 20 EDSFF drives per node with capacities of 3.84 TB to 61.44 TB, enabling up to 1118 TB* of raw capacity per node, or up to 5 PB usable capacity in an 8-node cluster.

Secure

Anchored by HPE's silicon root of trust, the server firmware creates a fingerprint for the AMD Secure Processor that must be matched exactly before the server will boot. This makes the HPE ProLiant DL325 Gen11 Server an ideal platform for virtualized workloads such as software-defined compute, CDN, VDI, and secure edge apps that require balancing processor, memory, and network bandwidth.

Hybrid

Engineered for hybrid operations, HPE ProLiant DL325 Gen11 integrates seamlessly with HPE GreenLake for Compute Ops Management. This as-a-service platform delivers simplified, secure management from edge to cloud, enhancing agility and operational efficiency across your entire compute and storage landscape.

Proactive system management

Included with every DL325 Gen11 server, HPE iLO 6 enables secure configuration, monitoring, and updates from virtually anywhere. It gives IT teams greater visibility and control, reducing complexity and improving response times for managing large-scale AI infrastructure.

* 1118 TB of raw capacity per node =
18x 61.44 TB capacity + 2x 6.4 TB

⁴“60 world records for performance and efficiency with HPE ProLiant.” HPE, 2024

Record-breaking technology

HPE ProLiant Gen11 Systems hold **60 world records** for performance and efficiency, including **7 in AI and ML** with Intel® and AMD powered servers.⁴



HPE ProLiant DL325 Gen11 for WEKA specifications

Table 1. HPE Solutions with WEKA on HPE ProLiant DL325 Gen11 specifications for node quantities, CPUs, SSDs, networking, and more



Node type	HPE ProLiant DL325 Gen11 purpose-built CTO for WEKA	
Minimum node quantity	8	
Maximum node quantity	100s	
CPUs per node	1x AMD EPYC (Genoa) 9454P 48 Core 2.75 GHz 256 MB L3 Cache	
SSDs per node	14x PCIe Gen5 read intensive E3.S NVMe	
SSD drive size	7.68 TB	15.36 TB
8-node usable capacity	484 TB	968 TB
Networking per node	2x NVIDIA ConnectX-7 400 Gbs NDR IB Single-Port OSFP, PCIe5 x16 1x NVIDIA ConnectX-6 LX Dual Port 10/25 Gbs SFP28, OCP NIC 3.0	
Software	WEKA Data Platform	
Data protection	<ul style="list-style-type: none">• Distributed data protection• Drive virtual hot sparing• Error detection: End-to-end data protection• In-flight and at-rest data encryption	
Protocols	POSIX, NFS, SMB, S3, GPUDirect Storage	
Snapshots and clones	File system level, up to 24,576 snapshots	
System monitoring	Cloud-based monitoring and analytics for application tuning and remote support	
System management	HPE iLO, HPE GreenLake for Compute Ops Management	
Minimum 8-node WEKA cluster performance	<ul style="list-style-type: none">• Sequential read bandwidth up to 180 GB/s• Sequential write bandwidth up to 64 GB/s• Random 4 KB read bandwidth up to 4.6M IOPS• Random 4 KB write bandwidth up to 1.1M IOPS• 162 microseconds with 4 KB read latency• 116 microseconds with 4 KB write latency	
8-node nominal power	3200 Watts	

Learn more at

[HPE Solutions with WEKA](#)

