Storage designed for AI workloads

HPE Solutions with WEKA

AI and the evolving data workflow

Artificial intelligence (AI) has moved beyond specialized applications and is now driving new, expansive demands on data workflows. AI workloads require higher throughput, increased input/output operations per second (IOPS), enhanced metadata performance, low latency, and massive scalability. In the realm of AI, the management of data is pivotal—everything depends on how data is utilized. As the market rapidly shifts toward large language models (LLMs) and generative AI (GenAI), the focus has also moved toward handling colossal datasets, more complex models, and an expansion from millions to billions of model parameters.

Raw data streams—including text, images, audio, and video—feed enterprises from diverse applications, data sources, and locations. These data streams may need to be transformed in real time to either be stored for long-term training or immediately used in inferencing processes. This rapid evolution impacts AI workflows and highlights the need for an emphasis on data pipelining. It raises critical questions about data storage locations, compute resource deployment, and their interactions with storage systems. The central question becomes: How must storage systems evolve to meet the new demands introduced by AI workloads?

A new type of storage for AI

A critical challenge in AI adoption is that while customers are optimizing compute and network infrastructure, they often overlook the importance of also optimizing their storage infrastructure—the very backbone that feeds and stores the growing volume of unstructured data so instrumental to driving AI insights. Without advanced storage solutions, even the most powerful networks and compute systems can't fully achieve their full potential, leading to bottlenecks and inefficiencies in AI workflows.

HPE GreenLake

Introducing a new purpose-built system: HPE Solutions with WEKA

HPE Solutions with WEKA delivers the ultimate platform for AI workloads by combining WEKA's ultrafast performance, scalability, and seamless AI framework integration with high performance infrastructure of Hewlett Packard Enterprise. WEKA's distributed architecture ensures low-latency access to large datasets and faster training and inferencing times. HPE data storage servers provide the high read and write throughput performance needed for AI workloads, as well as the low latency required for inferencing. The latest purpose-built system using HPE ProLiant DL325 Gen11 Server offers a low-cost 1U 1P solution that delivers exceptional performance, balancing compute, memory, and network bandwidth at 1P economics.

Accelerating time to insights and results through high performance

HPE Solutions with WEKA is designed with all-flash technology, leveraging NVMe solid-state drives (EDSFF and SSD). WEKA software directly accesses the underlying flash media in its native 4 KB format, enabling both small files and large datasets to be processed at record speeds. This technology can deliver up to twice the data throughput and supports both InfiniBand (IB) and Ethernet networking with bandwidth up to 400 Gbps. It also supports NVIDIA®'s dual-protocol functionality, allowing full use of existing network infrastructure while accommodating future upgrades to enhance throughput and reduce latency.

Exabyte scale and improved economics for increased ROI

HPE Solutions with WEKA can start with a cluster as small as 184 TB (raw) and 104 TB (usable), and scale up to 512 PB of all-flash storage, or 14 EB when including object tiering in a single namespace. The HPE ProLiant DL325 Gen11 Server system can support up to 20 EDSFF drives per node, making it the highest-density HPE purpose-built system for WEKA. This capability allows you to create infrastructure tailored to handle peak workloads, scaling for performance and capacity growth as system requirements evolve. The solution supports expanding a global namespace by tiering to any S3-compatible object store—whether HPE provisioned, third-party, or public cloud—for cost-effective, massive-scale storage.

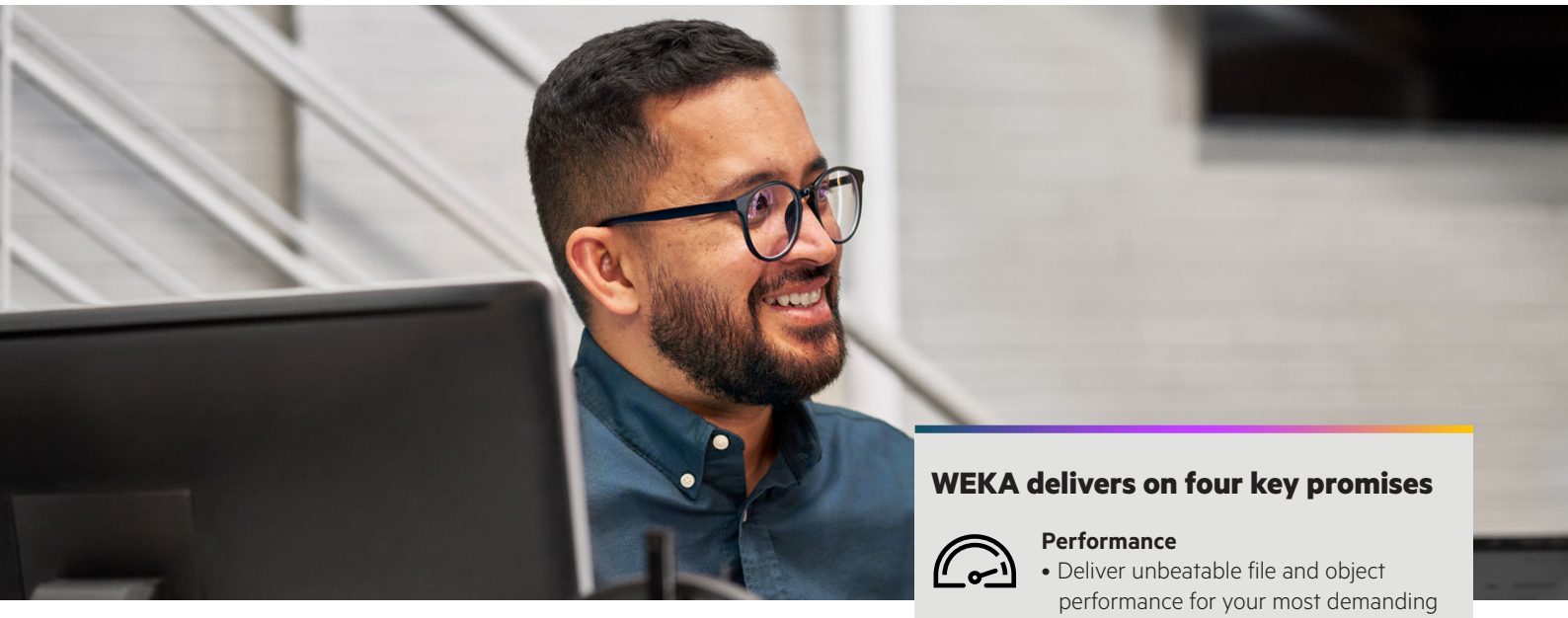
Designed for enhanced performance

One of the ways WEKA achieves exceptional performance is through the use of cutting-edge hardware technology from HPE, including the latest CPUs, high performance SSDs, and the fastest networking infrastructure. However, these same high performance

HPE ProLiant DL325 Gen11 Server hardware specifications

- **Processors:** Powered by 4th Generation AMD EPYC™ processors with 5 nm technology, supporting up to 128 cores, 384 MB of L3 cache, and 12 DIMMs for DDR5 memory up to 4800 MT/s
- **Memory:** 12 DIMM channels per processor for up to 3 TB of DDR5 memory, offering increased memory bandwidth and performance with lower power requirements
- **Data transfer:** Advanced data transfer rates from PCIe Gen5 with up to 2x16 PCIe Gen5, two OCP 3.0 slots, and EDSFF E3.S 1T drives
- **Networking:** Supports 1GbE, 10GbE, 25GbE, 100GbE, 200GbE, 400GbE, or IB, with support for the latest NVIDIA networking
- **Storage:** Available with 3.84 TB, 7.68 TB, and 15.36 TB NVMe drives, with up to 307 TB (raw) per node in the HPE ProLiant DL325 Gen11 Server system, supporting up to 20 E3.S drives per node. Total capacity with object store reaches up to 14 EB, and SSD capacity can go up to 512 PB
- **Expansion:** Storage is expanded by adding WEKA nodes or by adding additional drives to nodes. Clusters start with a minimum of eight nodes, with six drives/node
- **Form factor:** The HPE ProLiant DL325 Gen11 Server is a 1U 1P system, supporting up to 20 E3.S drives per node
- **Warranty:** Selectable levels of HPE Tech Care Service options available for HPE hardware
- **Recommended minimum** WEKA cluster starts at 8 nodes, 6 drives per node

components generate more heat than standard options, requiring careful design. A judicious balance of advanced components helps to fully utilize the WEKA Data Platform. While ala carte systems might seem cost-effective initially, they can sometimes lead to thermal issues and performance bottlenecks. HPE thermal engineers rigorously test and validate HPE configurations, at times making changes to what options are used to prevent potential thermal challenges while still helping ensure sustained high performance.



WEKA Data Platform

The WEKA® Data Platform is an award-winning, software-defined, high performance solution purpose-built for large-scale AI and analytics workloads. The WEKA Data Platform provides a single storage architecture that can run in the cloud, at the edge, on-premises, or in hybrid and multicloud deployments with the performance of all-flash arrays, the enterprise-grade feature set of network-attached storage (NAS), and the simplicity, scalability, and flexible economics of the cloud. It removes the barriers to data-driven innovation through an advanced software architecture optimized to solve complex data challenges and streamline the data pipelines that fuel GenAI, LLM, computer vision, and other performance-intensive workloads, delivering breakthrough performance and ease of use, simple scaling, and seamless sharing of data in virtually any location.

The WEKA Data Platform uniquely leverages the latest improvements in flash, networking, and compute to deliver the highest-performing AI/ML solution to customers.

WEKA delivers on four key promises

Performance

- Deliver unbeatable file and object performance for your most demanding applications supporting high I/O, low latency, small files, and mixed workloads with no tuning
- Accelerate training epochs and deliver faster inferencing

Massive scale

- Scale your compute and storage independently and linearly on-premises or in the cloud with WekaFS to handle 10s of millions or even billions of files of all data types and data sizes
- WEKA performance scales linearly as you scale capacity

Simplicity

- Helps eliminate the complexity and compromises of traditional data infrastructure with a single, easy-to-use data platform that helps eliminate storage silos across on-premises and the cloud

Sustainability

- Lower energy consumption and reduce the resulting carbon emissions by cutting data pipeline idle time, extending the usable life of your hardware, and moving workloads to the cloud

For WEKA software support

WEKA support

Learn more at

HPE Solutions with WEKA

Explore **HPE GreenLake**

Chat now (sales)

a00142745ENW

Hewlett Packard Enterprise