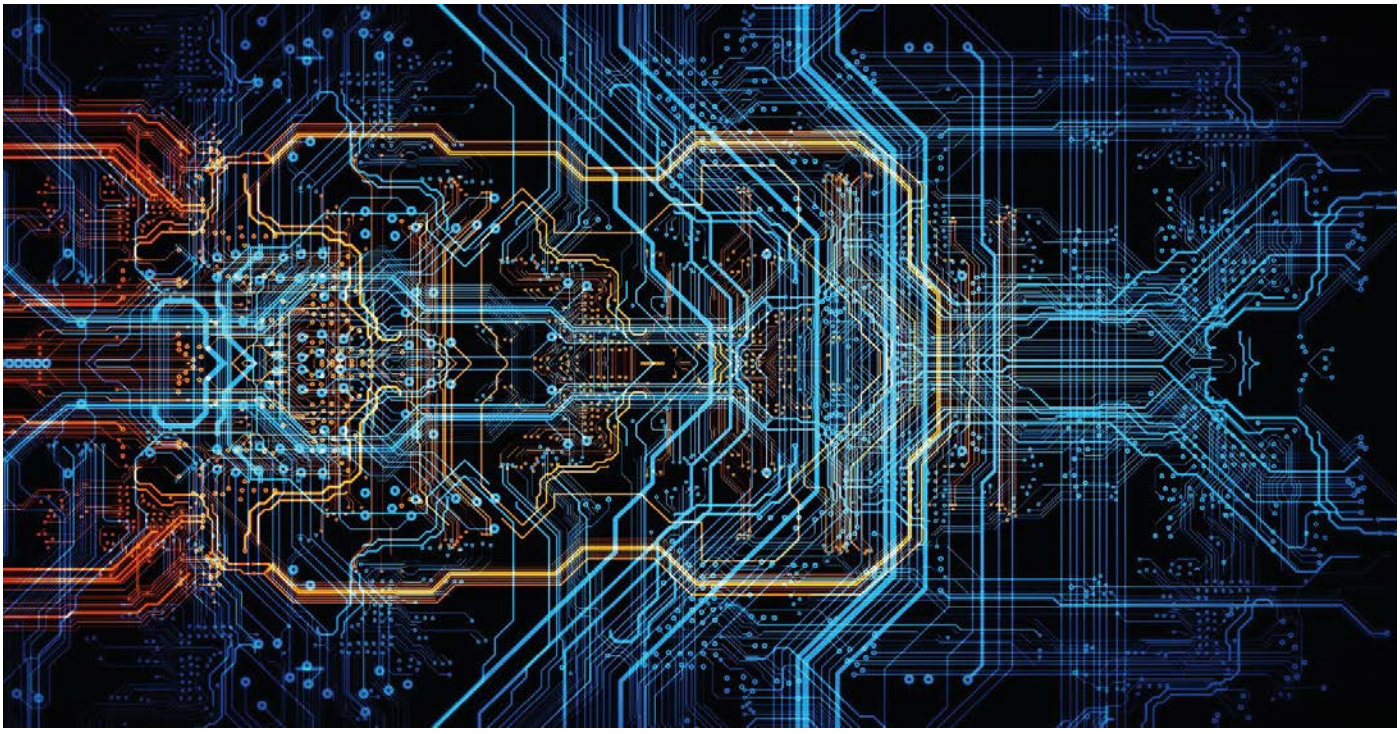
Hewlett Packard Enterprise

Reference Architecture

HPE Reference Configuration for HPE AI Data Node

Integrated two-tiered data storage for deep learning and high-performance computing



Contents

Optimizing storage for AI workloads

Artificial intelligence (AI) learning is moving from research into mainstream business use. In the Gartner 2019 CIO survey 37%[1] of respondents reported that their enterprises either had deployed AI or would do so shortly. Common examples of AI use are facial recognition, real-time translation from images, and voice recognition in cell phones. Many AI/machine learning workloads require storage solutions, which have been optimized both for working on very large data sets and for very high IOPS and/or throughput and low-latency performance. The expectation is that AI compute will come to resemble high-performance computing (HPC) in that not only will servers scale up, that is, adding more GPUs per server, but also scale out, that is, using a distributed clustered server environment. This will require the use of shared storage file systems to avoid storage bottlenecks.

Flash storage technology may be utilized to provide the necessary throughput performance but can be quite costly for capacity storage. As companies go into production with AI, data sets will grow to tens and even hundreds of petabytes, and will exceed the capacity of traditional storage appliances. To achieve scalability and performance while simultaneously controlling costs, storage system designers build separate tiers of storage for hot and cold data, utilizing archival object storage for the colder data. This dramatically lowers the total cost of ownership.

Hewlett Packard Enterprise, in partnership with WekaIO and Scality, provides storage solutions tailored to HPC and AI workloads using software-defined storage applications deployed on HPE ProLiant and HPE Apollo servers. With these solutions, customers can have high-performance, petabyte-scale storage solutions with integrated data lifecycle management, providing tiering management by file system and a single namespace. This solution can be implemented in a classic two-tier architecture, with one tier dedicated to high-performance flash while a second tier provides scalable object storage, typically as two separate clusters of storage servers. A second hybrid approach combines both tier elements into a scalable cluster, utilizing storage servers, which are optimized for both NVMe flash capacity and scale-out bulk data storage. This is the concept behind the HPE AI Data Node, based on the HPE Apollo 4200 Gen10 storage server. HPE AI Data Node offers a building block for production AI that can scale in performance and capacity.

What is HPE AI Data Node?

HPE AI Data Node is a reference configuration based on HPE Apollo 4200 Gen10, WekaIO Matrix™ parallel file system, and Scality RING object storage. HPE AI Data Node consists of:

- A storage-optimized HPE Apollo 4200 Gen10 server, scalable in clusters, provisioned with NVMe storage and capacity HDD on a 100GbE fabric
- WekaIO Matrix software, a high-performance parallel file system utilizing NVMe storage
- Scality RING, a scalable object store utilizing solid state and HDD high-capacity disks

With the data lifecycle management features built into WekaIO Matrix, colder data elements are automatically identified and tiered to S3-compatible Scality RING object storage. The entire data set is protected with distributed data protection scheme across a cluster of servers. This hybrid solution is a full-function high performance AI file store with an integrated and durable low-cost object storage tier, offering savings of up to half the infrastructure and operational costs of traditional solutions that deploy two separate storage clusters.

The HPE solution is based on HPE Apollo 4200 Gen10 storage server, which has been designed to simultaneously support large capacities of both NVMe and HDD, enabling it to be the converged platform in a hybrid AI data storage solution.

[1] "Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form," Gartner Inc., 2019

AI storage-tiered architecture for data lifecycle management

When designing high performance storage solutions, data movement tools are commonly employed to move data to the storage that provides the optimal cost/performance ratio for that data. WekaIO Matrix parallel file system has integrated this functionality, automatically moving colder data to lower cost tiers of object storage tier.

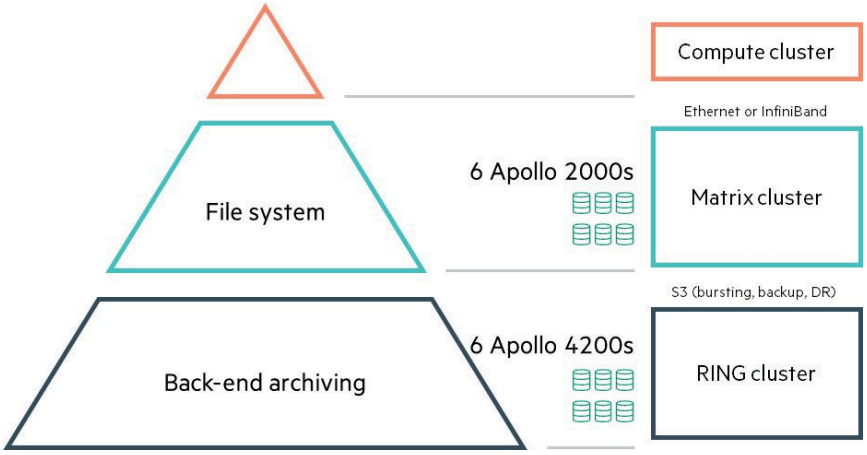


Figure 1. Disaggregated architecture for tiered AI storage

A disaggregated architecture for tiered AI storage, as shown in Figure 1, leverages Matrix functionality to present a single namespace across two separate data storage infrastructures. For this architecture, HPE recommends the performance tier be built on a cluster of HPE Apollo 2000 servers for the high-performance file tier and a second cluster built on HPE Apollo 4000 servers for the scale-out object tier.

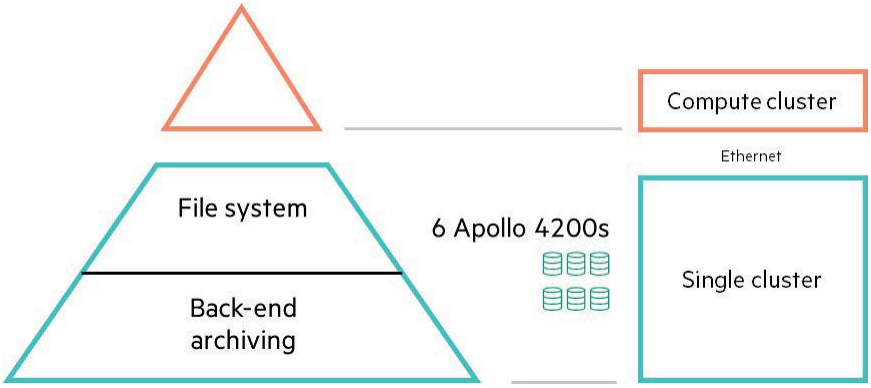


Figure 2. Hybrid architecture for tiered AI storage

A hybrid architecture, as shown in Figure 2, provides administrative benefits by combining the file system and back-end archiving tier into one simplified infrastructure, which contains both NVMe flash storage and HDD storage. The hybrid architecture operates in the same way as the disaggregated architecture. The two logical tiers, while physically combined into one cluster of server nodes, operates as two separate storage tiers running both the WekaIO Matrix performance file system and the Scality RING object storage tier. All data in both tiers is distributed in shards across the entire server cluster for performance, with distributed data protection to provide durability and availability even in the event of an unexpected server fault. The hybrid architecture preserves the data durability and data lifecycle management of the disaggregated model while not requiring any modification to the WekaIO or Scality software.

The challenge of such a hybrid architecture is finding a storage server that can be configured with sufficient capacities of both NVMe flash and HDD storage, to build the desired file-to-archive ratios. The HPE Apollo 4200 Gen10 meets these requirements with configurable options holding up to 46 TB of raw NVMe storage, and at the same time, up to 288 TB raw HDD storage, all contained in a standard 2U rack form factor. With the

high performance I/O of HPE Apollo 4200 Gen10, it is now possible to run both the high-performance file system and the scalable object storage software on one storage cluster. This results in operational savings from having half as many server nodes, network ports, rack space, power, and cooling as a distributed solution. The hybrid solution, when deployed on HPE Apollo 4200 Gen10, is referred to as HPE AI Data Node.

Recommended workloads for HPE AI Data Node

The following use cases are recommended for HPE AI Data Node storage.

- Machine learning—training and inference—autonomous vehicle

> "WekaIO Matrix was the clear choice for our on-premises deep neural network training… a NAS solution would not be able to scale to the extent we would need it to… and Matrix was the most performant of all the parallel file systems we evaluated."[2]
>
> - Dr. Xiaodi Hou, co-founder and CTO, TuSimple

- High I/O performance at extreme scale

> "TGen is dedicated to the next revolution in precision medicine—with the goal of better patient outcomes driving our core principles. Future-thinking companies like WekaIO, complement our core principle of accelerating research and discovery. The ability to run more concurrent high-performance genomic workloads will significantly advance our time to discovery."[3]
>
> - Nelson Kick, manager of HPC operations, TGen

The following are the industry examples where machine learning with massive data sets and requirements for extreme I/O storage performance is being utilized to solve business problems:

- Life Sciences, genomics and bioimaging
- Automotive (autonomous car programs)
- Oil and Gas (AI research)

HPE AI Data Node is recommended for data sets starting at a petabyte of total usable storage (combined hot and cold tiers) at hot-to-cold tier ratios up to 1:8, using Ethernet fabrics. HPE AI Data Node is not recommended for use as a general business file store where snapshot, dedupe, or incremental backup is required.

HPE AI Data Node infrastructure

WekaIO Matrix

WekaIO Matrix is the fastest, most scalable, parallel file system for AI and technical compute workloads that ensures applications never wait for data. WekaIO offers an NVMe-native, POSIX compliant file system that is fully coherent and resilient. The solution delivers the highest bandwidth, lowest latency performance to any InfiniBand or Ethernet-enabled GPU or CPU-based cluster.

To minimize idle time for compute clients, HPE partners with WekaIO for its high-performance shared storage. WekaIO Matrix includes the MatrixFS flash-optimized parallel file system, qualified on industry-leading HPE Apollo 4200 Gen10 servers. Matrix is a radically simple storage solution that delivers the performance of all-flash arrays with the scalability and economics of the cloud. Matrix transforms NVMe-based flash

[2] weka.io/solutions/enterprise/
[3] weka.io/solutions/research/

storage, compute nodes, and interconnect fabrics into a high-performance, scale-out parallel storage system that is well suited for I/O-bound use cases. WekaIO Matrix also provides automatic tiering and transparent migration of your cold data to Scality RING object storage to provide low cost and limitless scale.

WekaIO MatrixFS meets or exceeds the requirements of AI architectures. It is purpose-built with distributed data and metadata support to avoid hotspots or bottlenecks encountered by traditional scale-out storage solutions, surpassing the performance capabilities of even local NVMe storage. It supports distributed data protection (MatrixDDP) for data resiliency with minimal overhead and reliability that increases as the storage cluster scales.

Scality RING

Scality RING software-defined storage enables petabyte-scale, data-rich object storage services. Deployed on HPE Apollo 4200 Gen10 servers, it scales easily and infinitely with a mix of hardware, so as hardware evolves, adding capacity is easy, and RING takes advantage of server and media innovation over time. Acting as a single, distributed system, the RING scales linearly across thousands of servers, multiple sites, and an unlimited number of objects. It protects data with policy-based replication, erasure coding, and geo-distribution, achieving up to 14 nines of durability and 100% availability.[4] Regarded as one of the leaders in file and object storage by both IDC[5] and Gartner,[6] Scality RING supports native file, object, and AWS IAM and S3 interfaces, providing high performance across a variety of workloads at up to 90% lower TCO than legacy storage. HPE partners with Scality to provide a complete set of object storage configurations within HPE Scalable Object Storage with Scality RING, which supports both HPE Apollo 4200 and Apollo 4510 servers.

Along with that low TCO, RING offers superior performance over legacy storage and object-based systems. It ensures high throughput and low latency across small and large objects through its unique any-to-any performance capabilities. The platform's access and storage layers can scale independently from as few as three to thousands of servers.

The combined solution brings the best of market-leading Scality RING file and object storage on market-leading hardware from HPE for a best-of-breed solution with appliance-like experience, without appliance-like restrictions.

- **Economy and predictable costs:** Scality RING is the only storage solution to accommodate multiple workloads and lower TCO by allowing a mix and match of standard servers. Unlike conventional storage, Scality RING enables worry-free capacity expansion, upgrade, and swap out as data is managed by the software and not tied to appliance form factors.
- **Online:** Unlike data archived to tape, Scality RING keeps data online and available, and keeps the data intact, with up to 14 nines durability, including multisite options to tolerate entire site failure. Unlike other object storage, it also enables different durability and overhead ratios to match data value.
- **Scale:** Grows easily, cost-effectively, and without limits as stores of valuable data grow.
- **Compatibility:** More than 50 ISV partners, native support for object storage, and S3 for broad compatibility.

[4] scality.com/why-scality/
[5] "IDC MarketScape: Worldwide Object-Based Storage 2018 Vendor Assessment," IDC, 2018
[6] "Magic Quadrant for Distributed File Systems and Object Storage," Gartner, Inc., 2018

HPE Apollo 4200 Gen10 Server

HPE Apollo 4000 systems are purpose-built for large-scale deployments of software-defined object and clustered storage, analytics, or active archives. With HPE Apollo storage systems, companies harness Big Data and overcome data center challenges with optimized platforms that help unlock business insights and store data efficiently. HPE Apollo storage has been the platform of choice for many of the largest global 500 customers. HPE Apollo 4200 Gen10 Server delivers high-density storage with hundreds of terabytes of capacity in a 2U rack form factor. HPE Apollo 4200 Gen10, in an easily serviceable 2U design with up to 28 LFF or 54 SFF hot-plug drives, drives accelerated performance with a balanced architecture and NVMe connected SSDs. The Gen10 also features HPE iLO 5 and HPE silicon root of trust technology for firmware protection, malware detection, and firmware recovery.

Figure 3. HPE Apollo 4200 Gen10 server—up to 24 LFF hot-plug drives in a front-accessible expandable drive cage

The combination of NVMe connected SSDs and 24 LFF HDD capacity makes HPE Apollo 4200 Gen10 ideal for the HPE AI Data Node solution. All components of the solution can be purchased from HPE, offer HPE Pointnext multivendor support, and have been tested by HPE engineering.

Deployment guide for HPE AI Data Node

HPE has validated the capabilities of this solution as described in this paper. The solution is shipped as a bare-metal server. Once the systems are configured with an OS and attached to the customer's network, HPE partners will deploy their storage software products. The solution is supported by HPE Pointnext with collaborative multivendor support.

These steps are to be performed after the systems are shipped to the customer's site:

1. Rack the equipment; physically attach to the network
2. Create a VM on separate equipment for the Scality RING supervisor
3. Install Linux® OS (Red Hat® Enterprise Linux [RHEL] or CentOS v7.5 are the supported OS choices for this Reference Architecture.)
4. Set up networking for the cluster
5. Schedule WekaIO Matrix installation
6. Schedule Scality RING installation

Sample BOM

The sample BOM described in this paper builds a 6-node storage cluster. Two configurations are offered for different ratios of hot to cold storage. This BOM may be scaled to any desired capacity of storage, starting at six server nodes and growing in increments of three server nodes. To customize this BOM, contact your HPE HPC or storage solution architect.

- **Configuration 1**—1:6.7 ratio of hot to cold storage
  - Utilizing 7.68 TB NVMe SSD devices, minimum storage, with six server nodes:
    - A total of 132.7 TB of usable Matrix file storage
    - A total of 866 TB of usable RING object storage

- **Configuration 2**—1:8 ratio of hot to cold storage
  - Utilizing 6.4 TB NVMe SSD devices, minimum storage, with six server nodes:
    - A total of 110.6 TB of usable Matrix file storage
    - A total of 866 TB of usable RING object storage

Table 1. Sample HPE Apollo 4200 Gen10 BOM, single node (purchase 6 nodes)

Component	Part number	Quantity
HPE Apollo 4200 Gen10 24LFF CTO Svr	P07244-B21	1
HPE Apollo 4200 Gen10 6SFF NVMe Rear Cage	P07250-B21	1
HPE NVMe CPU2 x6 FIO Controller Mode for Rear Storage	P09657-B21	1
HPE XL420 Gen10 Intel® Xeon®-G 6132 FIO Kit	P08050-L21	1
HPE XL420 Gen10 Xeon-G 6132 Kit	P08050-B21	1
HPE 32GB (1x32GB) Dual Rank x4 DDR4-2666 CAS-19-19-19 Registered Smart Memory Kit	815100-B21	12
HPE Smart Array P408i-a SR G10 LH Ctrlr	869081-B21	1
HPE 96W Smart Storage Battery 260mm Cbl	P01367-B21	1
HPE InfiniBand EDR/Ethernet 100Gb 2-port 841QSFP28 Adapter	872726-B21	2
HPE X240 100G QSFP28 to QSFP28 1m Direct Attach Copper Cable	JL271A	2 or 4
HPE 1TB SATA 7.2K LFF LP DS HDD	861686-B21	2
HPE 12TB SATA 7.2K LFF LP 512e DS HDD	881787-B21	20
HPE 800GB SAS 12G Mixed Use LFF (3.5in) LPC SSD	P04531-B21	2
HPE 800W FS Plat Ht Plg LH Pwr Sply Kit	865414-B21	2
HPE Apollo 4200 Hardware Rail Kit	822731-B21	1
Configuration 1		
HPE 7.68TB NVMe x4 Lanes Read Intensive SFF (2.5in) SCN SSD	P10218-B21	6
Configuration 2		
HPE 6.4TB NVMe x4 MU SFF SCN DS SSD	P10226-B21	6

Note

HPE racks, networking equipment, and racking installation services are optional.

HPE Foundation Care or higher level must be purchased for this hardware to enable HPE Pointnext collaborative multivendor support.

HPE iLO standard features are supported under the **Server Hardware Warranty**. An HPE iLO Advanced or HPE iLO Advanced Premium Security Edition license is recommended. See support.hpe.com/hpsc/doc/public/display?docId=c04951959 for more information about HPE iLO licensing.

Table 2. Sample WekaIO Matrix BOM

Component	Part number	Quantity
Configuration 1		
WekaIO Matrix 1yr Subscription/Support per TB E-LTU for HPE Servers	Q9Q94AAE	277
WekaIO Matrix 1yr Tiering per TB E-LTU for HPE Servers	Q9R02AAE	609
Configuration 2		
WekaIO Matrix 1yr Subscription/Support per TB E-LTU for HPE Servers	Q9Q94AAE	231
WekaIO Matrix 1yr Tiering per TB E-LTU for HPE Servers	Q9R02AAE	655

Table 3. Sample Scality RING BOM

Component	Part number	Quantity
Scality RING Sgl Site 200TB HW LT E-LTU	P8Y90AAE	886
Scality RING Install Pkg 3 GS E-LTU	P8Y95AAE	1
Scality RING 24/7 Maint Single Site E-LTU	P8Z01AAE	886

Note

HPE AI Data Node requires a single site deployment of Scality RING, with ARC encoding (8/4) for highest data durability.

Summary

HPE has designed the storage-optimized HPE Apollo 4000 systems to be used in a wide range of Big Data analytics, software-defined storage, backup and archive, and other data storage-intensive applications. The HPE Apollo 4200 Gen10 architecture combines NVMe-connected SSD storage with HDD bulk capacity, enabling new ways to build software-defined storage solutions with integrated data lifecycle management in an ultra-dense manner to maximize data center efficiency.

The HPE AI Data Node provides a building block for production AI storage, with the performance and capacity to scale and grow with real-world operations. HPE AI Data Node leverages the storage performance of HPE Apollo 4200 Gen10 along with the performance of WekaIO and the capacity and efficiency of Scality RING to provide a foundational storage building block for production AI scenarios.

Resources and additional links

WekaIO Matrix

- WekaIO Matrix for HPE Servers QuickSpecs
- Accelerate time to value and AI insights technical white paper
- Accelerating AI capabilities with advanced storage solutions technical white paper
- Architecture guide for HPE servers and WekaIO Matrix
- Performance addendum for HPE servers and WekaIO Matrix technical white paper

Scality RING

- HPE Scalable Object Storage with Scality RING QuickSpecs
- HPE Scalable Object Storage with Scality RING on HPE Apollo 4200 Gen10 technical white paper

HPE Apollo 4200

- HPE Apollo 4200 Gen10 Walkthrough (YouTube)
- HPE Apollo 4200 Gen10 use cases ChalkTalk (YouTube)

Learn more at

hpe.com/storage/scalableobject
hpe.com/storage/apollo
hpe.com/storage/wekaio

a00065979enw, April 2019