Cerebras Storage Vendor Assessment for Distributed Cluster Workloads (concise)

Field	Value
Author	Eva Winterschön
Section	research/vendor-assessment (concise)
Version	0.4.1
Date	2025-08-08
Repo	https://github.com/evaw-cerebras/
Summarized	HPC, Cluster Storage, Performance Benchmarking
Aggregates	Docs + Reqs + Components (June+July 2025), RAG Analysis
Inferenced	Qwen3-235B-Instruct

Introduction

This analysis focuses on flash-based Network Attached Storage (NAS) systems that support **NFS over RDMA** (for low-latency POSIX file access) and **S3 object** protocols – a critical combination for HPC workloads that require both high-speed file I/O and scalable data lakes for unstructured data.

The key evaluation dimensions are **Performance**, **Security/Compliance** (primary weight), and **Total Cost of Ownership (TCO)** and **Scalability** (secondary weight).

Considerations on Assessing Storage for Cerebras Clusters

Due to the inherent design differences between Cerebras Systems 'Wafer Scale' approach to GPU-centric Ai/ML infrastructure, from physical hardware to software integrations, there are no direct comparisons yet possible when looking at existing storage vendor's solutions for Ai/ML workloads. We can approximate certain workload properties, baseline requirements for performance, and then evaluate the systems once we have an applicable integration.

Though we cannot directly compare Cerebras vs Nvidia *(and its common connectors like vLLM)*, we still need to analyze the potential solutions from all available sides of comparative analysis which provide insight into the potential of each product. Therefore, in this document there are references to vLLM and Nvidia and GPUDirect and so forth. These are included for the aforementioned reason.

Standardized Product Evaluation Workflow

A **standard evaluation workflow** ensures a fair and comprehensive assessment of each storage product. The process combines product testing with verification of vendor capabilities and compliance.

```
flowchart TD
        A[Define Requirements]<-->B[Identify Candidate Vendors]
        B --> C[Lab PoC: Performance Testing]
        B --> D[Lab PoC: Security & Compliance Testing]
        C & D --> E[Analyze Scalability & TCO]
        E --> F[Score & Rank Solutions]
        F --> G[Procurement Decision]
```

Define Use Case & Requirements

As these assessments are not specific to one infrastructure deployment, or one cluster client, we need to assess the solutions based on variable options and ranges of applicability. We are not necessarily settling on a single universal vendor in this exercise.

Stages of Defintion

- Gather specific workload requirements and constraints.
- Identify primary use cases (e.g. AI/ML training data, scientific simulations, virtualization datastores) and their I/O patterns.
- Determine needed protocols (e.g. NFS for POSIX file access, with RDMA for low latency; S3 for object access and data lake integration) and target performance metrics (throughput in GB/s, IOPS, and latency).
- Capacity needs (scaling from ~100 TB up to 10+ PB) and whether multi-tenancy or multi-site replication is required.

Identify Candidate Vendors

Shortlist vendors known for high-performance file/object storage.

- **VAST Data, Universal Storage**
- **Pure Storage, FlashBlade**
- **Oracle, ZFS Appliance**
- **TrueNAS, OSS ZFS storage**
- **Qumulo, scale-out NAS**
- **Scality, RING**
- **WekaIO, WekaFS**

Lab Testing & Evaluation Phases

Lab Proof-of-Concept (PoC): Performance Testing

Set up a test cluster (or use vendor provided test results) to measure performance under realistic conditions. Use a **reference cluster scale** (for example, start at ~100 TB and scale up to multi-PB) to verify that throughput and latency do not degrade at scale. Conduct single-node tests (one client node driving I/O to measure maximum per-client performance) and distributed tests (multiple clients in parallel to measure aggregate throughput and IOPS). Tools like **fio** (for low-level I/O patterns), **IOR** (for MPI-parallel large file throughput), and **mdtest** (for metadata operations like file creation) should be used to simulate real workloads. For example, run fio with random 4K reads/writes to simulate virtualization IOPS, run IOR with large sequential writes/reads for AI data streaming, and mdtest for file create/stat rates. Document the achieved throughput (GB/s), IOPS, and metadata ops, and ensure these meet requirements at 100 TB and at 1 PB+ with **no significant performance drop-off**.

Lab PoC: Security & Compliance Testing

Evaluate each product's security features in a controlled environment. Enable encryption (at-rest and in-transit) and measure any performance impact. Test access controls: set up NFSv4.1 or SMB shares with **ACLs** and verify that user and group permissions are enforced correctly across protocols. Test audit logging by generating file access events and ensuring they are recorded (and forwarded to a SIEM, if applicable). If possible, enable **compliance modes** (e.g. WORM/immutable snapshots for regulatory retention) and attempt tamper scenarios. Verify integration with security frameworks like LDAP/AD (for identity) and KMIP (for external key management). This step ensures the product can meet standards like **FIPS 140-3 encryption**, **NIST SP 800-53 controls**, and **ISO 27001** policies if required.

Analyze Scalability & TCO

Examine the architecture for scalability bottlenecks. Check maximum cluster size (number of nodes, total capacity supported) and whether performance scales linearly with capacity and client count. Key architectural features to note are: support for **RDMA networks** (InfiniBand or RoCE for low latency clustering), use of a distributed or centralized metadata design, and any limits on namespace (e.g. max number of files or objects). Also evaluate **TCO** by considering capital costs (price per TB of the solution at the scale needed) and operational costs (power/cooling, maintenance). Include any software license fees, required proprietary hardware, or extra networking costs (for example, solutions that *require* an InfiniBand fabric might increase cost). Some vendors use data reduction (compression, deduplication) and high-density flash (like QLC NAND) to lower $/TB – these factors should be included in TCO calculation. Normalize costs to a per-terabyte basis at ~1 PB scale for comparison.

Score & Rank Solutions

For each vendor, compile the results from performance, security/compliance, scalability, and cost analysis. Assign scores (e.g. 1–10) for each category. We use **weighted scoring** – in this scenario, **Security/Compliance** and **Performance** are primary (higher weight) considerations, reflecting a CISO's priority on security and an HPC architect's priority on performance. **TCO** and **Scalability** are secondary factors but still included. Calculate a weighted total score for each solution (details in a later section). This quantitative ranking, combined with qualitative findings (e.g. specific feature advantages or gaps), will form the basis of the procurement recommendation.

Procurement Decision

Finally, use the scorecard and organizational priorities to make a decision. If one solution leads in security and performance but has slightly higher cost, the organization must decide if that premium is justified. The final decision phase may involve negotiations, pilot deployments, and ensuring that the chosen vendor meets any additional requirements (such as support SLAs, data governance policies, etc.). At this stage, **legal and compliance checks** (e.g. verifying the vendor's certifications like FedRAMP, HIPAA, or supply chain considerations) are also completed before signing a contract.

Reference Cluster Scale and Performance Expectations

Throughout our evaluation, we consider a **reference cluster scaling from 100 TB to 10 PB** of usable capacity.

All solutions are expected to maintain **near-linear performance scaling** from smaller to larger capacities – i.e., a 1 PB deployment should sustain roughly 10× the workload of a 100 TB deployment (assuming proportionally scaled-out hardware) without performance **bottlenecks** emerging. The storage platforms are assumed to be deployed in a **private colocation data center** with high-speed networking (at least 100 Gbps Ethernet or EDR/HDR InfiniBand where applicable). Latency-sensitive HPC workloads (model training that uses NFS) benefit from Remote Direct Memory Access; thus, **NFS over RDMA** capability is an important criterion. Similarly, for big data workflows, an integrated **S3** object store interface is expected to handle analytics jobs and archival within the same system.

At 10 PB scale, even small inefficiencies can greatly impact performance, so we expect the vendor architectures to include features such as: parallel I/O handling, avoidance of single metadata server bottlenecks, and intelligent client load distribution. For example, the top platforms use techniques like distributing metadata across nodes or using client-side caching/parallelism to achieve high metadata operations per second. We also assume that the **performance is maintained under heavy concurrency** (hundreds of client nodes, or thousands of VMs/containers accessing storage). Our tests therefore stress multi-client concurrency (e.g. 128+ threads in IOR/mdtest) to ensure that throughput and IOPS scale with the number of clients without significant "tail latency" issues.

Key Evaluation Criteria and Weighting

We evaluate the vendors on four main criteria, with **Security/Compliance** and **Performance** as the primary weighted categories, and **Scalability** and **TCO** as secondary.

Performance

Measures how well the storage delivers I/O throughput, IOPS, and low latency under HPC workloads. This includes sequential read/write bandwidth (for big data and AI), random I/O performance (for virtualization or AI metadata reads), and metadata operation speed (file create/delete, directory listing under load). We also consider support for advanced performance features like direct GPU access (e.g. NVIDIA GPUDirect Storage) and how the system handles small files vs large files. A score of 10 would indicate top-tier, record-setting performance in independent benchmarks, whereas lower scores indicate performance trade-offs or bottlenecks.

Security & Compliance

Assesses the platform's security features, encryption, and compliance with industry and government standards. This includes encryption of data at rest (and whether it's certified to **FIPS 140-3** standards)[10][11], encryption of data in transit (end-to-end or via protocols like Kerberos for NFS[12]), access control capabilities (support for **ACLs** such as NFSv4 ACLs or Windows NTFS ACLs[13], role-based access control for management, multi-tenancy isolation), audit logging of user activities, and any special compliance modes (e.g. SEC 17a-4 WORM retention support, STIG hardening profiles, etc.). We also verify alignment with frameworks like **NIST 800-53** controls[14] and certifications such as **ISO 27001** (if the vendor has them). Given a CISO perspective, this category has the highest weight. A score of 10 means the product provides comprehensive security features with relevant certifications (e.g. FIPS-validated encryption, built-in audit trails, RBAC, and demonstrated compliance reports), whereas lower scores mean some gaps (for example, lack of audit logging or pending certifications).

Scalability

This criterion examines both *capacity scalability* (how many petabytes and file count it can grow to) and *performance scalability* (ability to add nodes to increase throughput linearly). It also considers global namespace capabilities (can it present a single namespace across all 10 PB or does it require multiple volumes?), support for geo-distribution or multi-site replication, and client scalability (number of concurrent client mounts supported, etc.). We check if the architecture is truly scale-out (e.g. adding more storage nodes adds performance) and features like **pNFS** or client parallelism that aid scaling. A top score indicates the system can handle **multi-petabyte, billions of files, hundreds of nodes** deployments with ease (examples: proven deployments at >50 PB, or design with no hard cluster size limit), whereas a lower score might indicate a more scale-up design or practical limits (e.g. max 2 controllers, or degrading performance beyond a certain cluster size).

Total Cost of Ownership (TCO)

We evaluate relative cost-effectiveness, including upfront acquisition cost, licensing, and operational costs over 3-5 years. Lower $/TB (especially at scale) and efficiency (power/cooling per TB) result in higher scores. We also consider features that improve TCO: inline data reduction (which effectively lowers cost/TB stored), use of commodity components vs proprietary hardware, and the ability to start small and scale without painful forklift upgrades. Furthermore, ease of management (which affects OpEx) can factor in. Since exact pricing is often vendor-specific and negotiated, we score this qualitatively, assuming typical market prices. An open-source or commodity-based solution like TrueNAS or Scality might score very high on TCO (due to low software costs and flexibility), whereas a proprietary all-flash appliance might score lower unless it demonstrates significantly better density or data reduction that justifies its cost.

Weighting and Why and How

Using a normalized scoring process for each priority layer, and accordingly assign percentages.

- **30%** to *Performance*
- **30%** to *Security/Compliance*
- **20%** to *Scalability*
- **20%** to *TCO*

This reflects the primary importance of security and performance in HPC storage selection, with acknowledgment that scalability and cost are secondary, though still significant. Each vendor will receive a numeric score in each category, multiplied by these weights to produce a weighted total.

Security Compliance Feature Matrix (Capabilities vs. Requirements)

Simplified Table

See the full document as well as our CISO checklist document for detailed aspects of security requirements.

Security Feature / Compliance	IBM Spectrum Scale	Qumulo Core	Scality RING	WekaFS
Encryption at-rest	Yes	Yes	Yes	Yes
Encryption in-transit	Yes	Yes	Yes	Yes
Access Control & Multi-tenancy	Yes	Yes	Yes	Yes
Audit Logging	Yes	Partial	Yes	Partial

Security Feature / Compliance	VAST Data	Pure FlashBlade	Oracle ZFS	TrueNAS Enterprise
Encryption at-rest	Yes	Yes	Yes	Yes
Encryption in-transit	Yes	Yes	Yes	Yes
Access Control & Multi-tenancy	Yes	Yes	Yes	Yes
Audit Logging	Yes	Partial	Yes	Improving

Security Compliances For detailed listing of each product's compliance certificates, please see the full version of this document.

Mapping to CISO priorities

CISO weights encryption (FIPS-certified) and audit logging very highly, followed by access control capabilities and compliance proofs. Thus, in our scoring, products like VAST, Pure, and Qumulo, which excel in encryption and auditing, score well on Security/Compliance, whereas those that require add-ons or lack certification yet (TrueNAS, Weka) score a bit lower despite having the essential capabilities.)*

Performance and Architectural Features

To assess performance, we consider both raw benchmark results and the architectural features that enable or limit performance. Below we note each solution's support for critical performance-related features and any known limitations:

Remote Direct Memory Access (RDMA) Support

RDMA bypasses TCP/IP overhead to deliver low-latency, high-throughput access. Several vendors support NFS over RDMA or similar:

- *VAST Data:* **Yes.** VAST supports NFS over RDMA (RoCE or IB) for both NFSv3 and NFSv4.1, enabling very low latency file access[57]. Internally, VAST's disaggregated shared-everything architecture uses NVMe-oF over 100 Gb Ethernet or InfiniBand between compute nodes and storage enclosures[58][59].
- *Pure FlashBlade:* **Yes.** FlashBlade supports NFS using RDMA (over converged Ethernet or InfiniBand) as an option for high-speed data transfer[4]. In tests, NFS with RDMA on FlashBlade achieved improved throughput and lower CPU usage on clients[4]. *(Pure is also introducing S3 over RDMA in new releases to accelerate object performance[60].)*
- *Oracle ZFS:* **Partially.** Oracle ZFS Appliance supports RDMA for block (iSCSI over InfiniBand and RoCE)[61], but NFS runs over TCP/IP (NFS over RDMA is not explicitly advertised; clients would typically use IP over IB if anything). The primary network for NFS is 100 Gb Ethernet or IB with IPoIB.
- *TrueNAS:* **Limited.** TrueNAS SCALE (Linux-based) can leverage RDMA for iSCSI (iSER) and potentially NFS with manual setup, but by default RDMA is off[62]. TrueNAS CORE (FreeBSD) does not support NFS RDMA. In practice, TrueNAS is usually deployed on standard Ethernet (TCP/IP).
- *IBM Spectrum Scale:* **Yes (cluster-internal).** Spectrum Scale (Storage Scale) uses RDMA for inter-node communication in the cluster for data and metadata access, supporting InfiniBand and RoCE fabrics for its NSD protocol[63][64]. However, when exporting Spectrum Scale via NFS (through CES protocol nodes), that NFS traffic is over TCP/IP (no RDMA for NFS exports)[65]. Many Spectrum Scale HPC deployments use the native client with RDMA instead of NFS.
- *Qumulo:* **No.** Qumulo is designed for Ethernet networks; it does not support NFS/RDMA. It uses TCP/IP over high-speed Ethernet (clients commonly use 40/100 GbE). The focus is on scale-out over IP; adding RDMA could be future but not currently in documentation.
- *Scality RING:* **No (for file access).** Scality's S3 object access is HTTP/TLS over Ethernet. The NFS/SMB connectors are essentially gateway servers that use TCP/IP to talk to clients. There's no RDMA for those connectors. The RING backend sync between nodes is over standard networking (Ethernet, can be 10/25/100 GbE).
- *WekaFS:* **Yes.** Weka was built to exploit RDMA. It supports InfiniBand and RoCE for the cluster fabric; Weka clients use a user-space RDMA-enabled protocol to talk to the Weka cluster, achieving very high IOPS and low latency (comparable to local NVMe). For compatibility, Weka also supports NFS (v3/v4) and SMB, which would be over TCP, but those are typically slower paths used only when necessary. Most HPC users run the Weka client over RDMA for maximum performance.

Parallel NFS (pNFS) / Protocol Parallelism

Instead of a single NFS server bottleneck, some systems allow parallel access:

- *VAST:* NFS ops are handled by scale-out **CNodes** (compute nodes). VAST allows mounting the cluster via multiple IPs and uses NFS multipath/multichannel and RDMA to parallelize traffic[2]. It's not standard pNFS, but the effect is similar (all CNodes can serve data in parallel since there's no single metadata master)[66][67].
- *Pure FlashBlade:* Each FlashBlade node serves a portion of the data; clients automatically distribute NFS requests across blades. While not using the pNFS protocol per se, FlashBlade's design stripes data and metadata across blades, achieving parallelism. It supports NFSv4.1 but relies on its own scale-out rather than client-driven pNFS.
- *Oracle ZFS:* Does **not** support pNFS. It is essentially a dual-controller system (active-active). Clients mount a single head IP. There is no parallel NFS across multiple nodes (since only 2 controllers). So, metadata and data funnel through those controllers (which are very powerful and have lots of DRAM cache to boost performance).
- *TrueNAS:* No pNFS; TrueNAS is also essentially a dual-controller (HA pair) system at best. No horizontal scale-out for a single share (except through external clustering like Gluster in TrueNAS SCALE which is different and not pNFS).
- *IBM Spectrum Scale:* Spectrum Scale's native client offers parallel access to data across many cluster nodes (with a distributed locking mechanism). If using NFS via CES, there is an NFSv4.1 implementation that can support pNFS layouts for data (Spectrum Scale can present an NFSv4.1/pNFS endpoint, though this is less common than using the native client). Generally, the native approach is superior: applications use GPFS client which stripes I/O across multiple NSD servers in parallel.

- *Qumulo:* No pNFS, but the system itself is cluster-aware. Any node can serve any data (Qumulo's file system distributes data and metadata across nodes). Clients typically mount via a virtual IP that is moved to the node owning the data or use round-robin DNS to multiple nodes. The experience is not pNFS, but the cluster can deliver parallel throughput when multiple clients connect to different nodes.
- *Scality:* The file access is through connectors (which are basically stateless NFS heads). Scality's architecture suggests you can load-balance NFS requests across multiple connector nodes, but each file operation goes through one connector at a time (no striping of a single large file via pNFS). S3 calls go to any RING node via a load balancer – object operations can scale out massively (since any node can handle object PUT/GET and internally data is spread).
- *Weka:* Weka doesn't use pNFS because it has its own highly parallel client protocol. It allows many nodes to concurrently read/write different parts of the filesystem with a global coherent view. With NFSv4.1 support, Weka acts as an NFS server (not pNFS). But Weka's strength is using its own client for parallelism, achieving extremely high aggregate throughput (many clients reading/writing in parallel scale linearly until hitting network limits).

Metadata Handling & Small File Performance

A frequent performance bottleneck in distributed storage is metadata (file creation, listing, small file I/O).
Notable approaches:

- *VAST:* No dedicated metadata servers – metadata is stored in byte-addressable storage (Optane/QLC) distributed and accessible by all CNodes[68][69]. This means metadata operations scale with the number of CNodes and are very fast (VAST has demonstrated millions of file creates/sec). Small files are also handled efficiently by writing them to byte-addressable storage class memory (Optane) before flushing to QLC flash.
- *Pure FlashBlade:* Uses a distributed metadata database on the blades[70]. Each blade contributes to metadata performance. FlashBlade was explicitly designed to handle **millions of small files** without special tuning[71]. It uses an internal key-value store on flash to track metadata, yielding high metadata ops and eliminating single points of serialization.
- *Oracle ZFS:* Metadata is handled by the controllers using DRAM and SSD cache. Oracle ZFS excels at streaming workloads; for *very small* file workloads its performance is good (due to DRAM cache) but not as scalable as the truly distributed systems. It can handle enterprise NAS duties (directories with millions of files) but the two-controller design limits the total metadata ops compared to scale-out solutions.
- *TrueNAS:* Similarly, TrueNAS (based on OpenZFS) caches metadata in RAM and L2ARC (SSD). Metadata ops are decent on a single system, but scaling them requires scale-up (more CPU/RAM in the box). TrueNAS isn't designed for billions of tiny files in one namespace at high ops rates – it's better for medium workloads or as storage behind other systems.
- *IBM Spectrum Scale:* Extremely strong in metadata scaling. Spectrum Scale can designate multiple metadata manager nodes and uses a distributed locking mechanism. It even allows separating directory subtrees to different metadata nodes if needed. In practice, Spectrum Scale has achieved high metadata performance in benchmarks (the record for metadata ops was often on GPFS). It also has features like **metadata replication** for reliability and can store metadata on faster tiers.
- *Qumulo:* Qumulo's file system was built with a metadata-first architecture. It maintains a real-time database of metadata and usage. Small file ops are a known strength – Qumulo has internal metrics counting every file, and they claim efficient handling of billions of files. Its distributed file system spreads metadata across nodes and uses flash (SSD) for metadata acceleration (with data on SSD/HDD hybrid or all-flash nodes).
- *Scality:* For object workloads, Scality stores metadata on a subset of nodes or across nodes depending on configuration. It's designed for billions of objects (metadata scalability is very high in object mode). For file (NFS/SMB), the connectors translate file ops to object ops – small file I/O may incur overhead (object conversion), so not as fast as pure file systems for metadata ops. However, in recent versions, Scality has **Metallica** (metadata accelerator) to speed up file metadata by storing it on SSD separate from object data.
- *Weka:* WekaFS keeps all metadata in the distributed memory of the cluster, with redundancy. Metadata operations are optimized via its internal distributed structure. Weka is known to handle small files very well (it avoids typical metadata bottlenecks by distributing the directory entries across nodes and using an efficient network protocol). Many AI workflows with lots of small files (like millions of images) have found Weka's performance to be strong.

Client Concurrency & Throughput

In HPC and AI clusters, it's common to have **hundreds of clients** hitting the storage simultaneously (e.g., 1000 GPU servers reading training data). We consider how each handles concurrency:

- *VAST:* Designed for "embarrassingly parallel" scale[72]. All clients can utilize all storage nodes thanks to the disaggregated shared-everything model. Locking is fine-grained and distributed, so concurrent reads/writes scale well. VAST specifically markets high aggregate bandwidth with many clients (they've published multi-client SPEC SFS benchmarks showing linear scaling).
- *Pure FlashBlade:* Each client connection is handled by a blade, and FlashBlade's network stack allows spreading clients across blades. It has a high number of network ports and can sustain many concurrent flows. In independent tests, FlashBlade demonstrated near-linear scaling to dozens of clients for read throughput. One limitation might be that extremely metadata-heavy concurrent workloads could contend on internal resources, but overall concurrency handling is robust.
- *Oracle ZFS:* With two controllers, there's a limit to how many simultaneous heavy clients it can handle before saturating CPU or network ports on those controllers. It can support hundreds of NFS mounts, but the total throughput is capped by the controllers (18 GB/s max throughput as per Oracle's tests[73]). For moderate concurrency, it performs very well (especially with DRAM cache absorbing IOPS), but at extreme thread counts, it won't match the linear scaling of a cluster solution.
- *TrueNAS:* Similar to Oracle, a single-node TrueNAS has finite CPU and network. It can handle dozens of concurrent clients, but 100+ heavy clients would tax the system. TrueNAS SCALE could be clustered with e.g. GlusterFS for concurrency, but then performance suffers. In essence, TrueNAS is better for small to mid-size concurrency scenarios.
- *IBM Spectrum Scale:* Excellent concurrency support. With many nodes in the cluster serving data, Spectrum Scale can handle very large numbers of clients. HPC centers use GPFS with thousands of compute nodes – the locking and caching algorithms are designed for that scale. Some tuning is needed for extreme cases (e.g. avoiding lock contention on single directories by spreading files), but it's field-proven in supercomputers.
- *Qumulo:* Built with concurrency in mind for media and large enterprise loads. All nodes share the load; Qumulo's distributed file system and intelligent client connection routing mean 100+ clients can be spread over 100+ cluster threads fairly evenly. Internal testing by Qumulo showed linear scaling to at least 50 nodes/clients. The system also has built-in analytics to show if any client is "hot" – which helps ensure one client doesn't starve others.
- *Scality:* For object, concurrency is a non-issue – it's what object stores do best (many clients PUT/GET in parallel). For NFS/SMB, the connector nodes could become chokepoints if too many clients hit a single connector. The solution is to deploy multiple connector VMs and load balance clients among them. This works, but it's a bit more manual to scale to hundreds of clients (e.g. you might run 10 NFS connector VMs to handle 100 clients, 10 each). The backend RING can more than handle the throughput as long as front-end connectors scale out.
- *Weka:* Concurrency is a strong suit. Weka's client I/O is asynchronous and parallel; many clients can hit the cluster and Weka will distribute the load across all storage nodes using a fast network fabric. In SPEC SFS tests and internal benchmarks, Weka has shown very high ops with multiple clients. Its architecture avoids single-server bottlenecks by not having a single choke point for data or metadata.

Namespace & Capacity Scale

All these solutions claim to scale to **petabyte** levels, but actual limits:

- *VAST:* Scales to **exabytes** in theory (they advertise supporting 10s of petabytes in a single namespace, and the architecture is designed for "universal scale"). Metadata is global, so you don't need multiple volumes. They've publicly mentioned deployments over 100 PB[74]. One cluster can have 10s to 100s of nodes. There is effectively one namespace (with tenants separated by directory and policies).
- *Pure FlashBlade:* The new FlashBlade//S can scale to multiple chassis. A single chassis has up to 10 Blades; and multiple chassis can be interconnected. Pure hasn't published a hard limit in 2025; earlier models scaled to ~150 blades. Let's assume it can handle at least tens of petabytes (they mention 8.8 PB in a full FlashBlade//S deployment in one namespace[75]). The namespace is unified across all blades in the cluster.
- *Oracle ZFS:* Max capacity ~8.8 PB (with the ZS11-2 using all-flash drives) in one system[75]. If more capacity is needed, you'd have multiple appliances and use a manual federation (not a single namespace across appliances). So, it's not meant for beyond ~10 PB in one unit.
- *TrueNAS:* Depending on model, a single TrueNAS can hold on the order of a few PB (using expansion shelves, maybe up to ~10 PB raw in a top-end configuration with many disk shelves). But again, it's one head unit (or HA pair) managing that. No single namespace beyond that except via external clustering (which is not mainstream for TrueNAS yet). So practical scale: low PBs per system.
- *IBM Spectrum Scale:* There are production Spectrum Scale systems in the **100+ PB** range (e.g., at large research labs). It can incorporate thousands of disks across many nodes. It also can tie into IBM Cloud Object Storage for tiering beyond that. The namespace can handle billions of files. It's essentially limited by hardware and some theoretical limits (which are very high, like 2^64 files).
- *Qumulo:* Qumulo's largest public references mention clusters in the ~10s of PB and billions of files. The architecture can scale to at least 100 nodes, possibly more (they soft-limit at around 100 nodes for tested configurations). Each node contributes capacity, so 100 nodes * ~200 TB/node = 20 PB usable, for example. Namespace is single and global.
- *Scality:* Scality RING can scale to **100s of PB** easily by adding standard servers. It's used by cloud providers and telcos for massive object stores. It's perhaps the most scalable in capacity (object storage nature). File system access via connectors doesn't change that capacity limit (it just might be unwieldy to have billions of files via NFS, but technically possible by splitting across buckets/exports). Scality has installations north of 200 PB in production (object data).
- *Weka:* Weka can scale to multiple racks of NVMe servers; known deployments are in the 10–30 PB range of usable data for high-performance storage. Weka is often used as a fast tier (with older data offloaded to object storage). The metadata design could support billions of files; they've demonstrated handling very large directories in tests. If more than ~30 PB is needed, one might deploy multiple Weka clusters (or use their new "N + 1" clustering concept), but within a single cluster, you might be limited by practical network and node counts (perhaps on the order of a few hundred NVMe devices aggregated).

Assessments **VAST, FlashBlade, Spectrum Scale, Qumulo, Scality, and Weka** all provide substantial performance and can handle at least 10 PB in one cluster, with varying degrees of linear scalability.

Efficiency at Certain Scale **Oracle ZFS and TrueNAS** are more limited in scale-out but can be very effective at smaller scales (and may be more cost-effective there).

Distributed Benefits The architectural notes highlight that the truly distributed systems (VAST, IBM, Weka, etc.) avoid single points of contention (using RDMA, distributed metadata, etc.), which typically translates to superior performance at scale – as reflected in our scoring.

Comparison Matrix and Scoring Considerations

The following tables summarize how each vendor's solution ranks on the primary evaluation criteria.

- **Table 1** presents a high-level feature/support comparison for quick reference
- **Table 2** provides the weighted scoring for each criterion (Performance, Security/Compliance, Scalability, TCO), with some rationale.

Key Features & Protocols Supported

Vendor	NFS over RDMA?	S3 Protocol Support	All-Flash Performance (at 1 PB)	Notable Security Certifications/Features
VAST Data	Yes (NFSv3 & v4.1 over RDMA)[57]	Yes (native S3 interface)[57]	Excellent – scale-out NVMe, >100 GB/s achievable (RDMA & NVMeoF)	FIPS 140-3 validated encryption[10]; comprehensive audit[42]; Mult
Pure FlashBlade	Yes (NFS over RDMA supported)[4]	Yes (native S3 interface)	Excellent – scale-out blades, ~10–20 GB/s per chassis[73]	FIPS 140-2 crypto[15] (140-3 in progress); Always-on encryption; Co
Oracle ZFS	Partial (RDMA for block, not NFS)[61]	Yes (native S3, OCI-compatible)[91]	Very Good – up to 18 GB/s per system[73] (limited scale-out)	Cohasset-certified WORM compliance[94]; FIPS 140-2 (Solaris)[93];
TrueNAS	Limited (iSER for iSCSI, NFS RDMA not official)[62]	Yes (via built-in MinIO S3 server)	Good – ~5–10 GB/s on high-end HA setup (scale-up only)	OpenZFS checksums for integrity; optional FIPS 140-3 module[100]
IBM Spectrum Scale	Yes (RDMA for GPFS client)[63][64]	Yes (via Object CES gateway)[118]	Excellent – 100+ GB/s with large cluster (virtually linear scale)	Encryption w/ external KMS[20]; File Audit Logging[46]; integrates v
Qumulo	No (Ethernet only)	Yes (native S3 API)[7]	Very Good – linear scale-out, e.g. 20–50 GB/s with cluster	FIPS 140-2 encryption certified[22]; robust file audit logs[35]; AD in
Scality RING	No (file via connectors on TCP)	Yes (native S3 & Swift)[8]	Good (object): high multi-stream throughput; Moderate (NFS): depends on connectors	Cohasset SEC 17a-4 compliance (WORM)[51]; IAM for S3 multi-tena
WekaFS	Yes (InfiniBand/RoCE for client)	Yes (native S3 front-end)[9]	Outstanding – extremely low latency, >100 GB/s in small cluster (GPU-optimized)	End-to-end encryption (client to disk)[25]; impending FIPS 140-3 ce

Considerations

1. All vendors support NFS and SMB core file protocols (SMB omitted for brevity; all support SMB except perhaps Weka which added it recently).
2. "All-Flash Performance" is a qualitative estimate for a 1 PB-ish system; actual performance will vary.
3. RDMA support is a key differentiator for latency: VAST, Pure, IBM (GPFS), Weka leverage it heavily, whereas others use optimized TCP. S3 support varies in maturity: Scality and VAST are very mature, Qumulo and Weka are newer but functional.

Weighted Solutions Scoring

Scores are on a 1–10 scale (10 = best) for each category, and the weighted total (out of 100) is computed using the weights:

- Performance 30%
- Security/Compliance 30%
- Scalability 20%
- TCO 20%

Scores are based on the detailed analysis above and available data points [119][120].

Vendor	Performance (30%)
VAST Data	9 – Exceptional throughput & low latency (RDMA, scale-out), excels in AI/ML[57]. Small file perf very strong.
Pure FlashBlade	9 – Excellent mixed I/O performance[70], easy to deploy. RDMA support improves throughput[4]. Limited mainly by cluster size (blades).
Oracle ZFS	7 – Very high performance for a dual-head system (lots of cache, 18 GB/s+ throughput)[73]. But cannot scale-out beyond 2 controllers – one system's limits.
TrueNAS	6 – Good performance at small scale; all-flash TrueNAS can saturate ~100 GbE. Lags in parallel scaling (single controller limits). Suitable for moderate workloads, not top-end HPC peaks.
IBM Spectrum Scale	9 – Excellent performance at scale (proven in top HPC sites). Nearly unlimited concurrent throughput when properly configured. Minor latency overhead for small ops prevents a perfect 10.
Qumulo	8 – Very good performance for majority of workloads. All-flash Qumulo clusters can rival Isilon/Pure in throughput. Slightly lower max performance than Weka/GPFS in ultra-scale or low latency due to no RDMA, but more than sufficient for most us
Scality RING	7 – Object performance: high aggregate throughput, but single-stream or single-directory file ops not as fast as others. NFS via connector adds latency and some bottleneck. Great for large sequential or parallel reads/writes, less so for metadata-h
WekaFS	10 – **Best-in-class** performance for demanding HPC/AI (orders of magnitude faster on small random and multi-gigabyte throughput per client)[114]. Takes full advantage of flash and GPU-direct I/O.

Considerations In the scoring methodology, a difference of 1 point is significant weight. For example, Performance: Weka at 10 sets the bar, VAST/Pure/IBM at 9 just shy, TrueNAS lowest at 6. Security: VAST perfect 10 due to complete feature set and validation[10][42], others 8-9 if minor gaps or not certified. Scalability: Scality and IBM at 10 for essentially unlimited scale, Weka lower not for performance but for capacity methodology. TCO: TrueNAS 10, Pure lowest 6 due to premium pricing, etc.)*

- **VAST Data** has the highest weighted score (91), excelling especially in security and performance, making it a top choice for HPC Flash NAS if budget permits and extreme security is needed.

- **Scality RING** (86) and **Spectrum Scale** (85) lead for environments requiring maximal scalability (Scality more for capacity, IBM for all-around HPC performance) – they bring strong security and reasonable cost at scale too.

- **Qumulo** (85) and **WekaFS** (85) also tie as leaders: Qumulo offers a great balanced solution (slightly easier path, broad capabilities), while Weka offers unparalleled performance (ideal for AI labs) with good security, only held back by slightly narrower scalability in capacity.

- **Pure FlashBlade** (83) is close behind – extremely polished and high-performing, just more costly and closed in scale.

- **Oracle ZFS** (74) and **TrueNAS** (72) score lower mainly due to scalability limits; however, they remain excellent choices in their niches (Oracle for integrated enterprise/cloud environments with compliance needs; TrueNAS for budget-conscious deployments that still need flash and decent security).

It's important to note that scoring is based on an assumed weighting; individual customer-facing organizational priorities could shift the outcome. For instance, if TCO is paramount, TrueNAS might rank higher despite lower performance, etc. Also, all these solutions are proven – lower score does not mean "inadequate," but rather "less optimal in this specific weighted evaluation."*

Simplified Analytical Benchmark Guide

The *Analytical Benchmark Guide* outlines the same evaluation process and criteria, emphasizing performance and security outcomes, and can be used internally to validate architectures minus any vendor cost considerations or legal terms.

Evaluation Workflow Summary (Technical Focus)

Define Technical Requirements

- Perf requirements: throughput, IOPS, latency
- Capacity requirements in deployment and scale-out
- Security requirements (online/offline/transit/etc)

Select Candidates & Architecture Planning Basing on requirements, plan a test deployment architecture for evaluations

Conduct Performance Benchmarks

- Deploy or obtain access to each candidate system in a test environment with sufficient load-generating systems to simulate the target environment.
- Ensure tests are run at different scales (at 10% capacity and near 80% capacity) to see if performance remains consistent as the system fills up.
- Document any tuning done (did the system require manual stripe settings, or did it self-optimize?).

Throughput tests Run benchmarks with multiple processes on client nodes reading/writing large files (tests sequential read/write in parallel). Note achieved aggregate GB/s and how it scales from 1 node to N nodes.

IOPS tests Run benchmarks with random 4K reads/writes from multiple clients to test random I/O performance (simulate virtualization or AI metadata access). Measure IOPS and latency distribution.

Metadata tests Run **mdtest** to create and stat large numbers of small files across the filesystem. Record operations/sec.

Evaluate Security Features For each system, verify the availability and functionality for our baselines:

Encryption: Enable encryption and ensure data is indeed encrypted (perhaps by trying to read disks directly or checking for encryption keys). Check if encryption can be enabled without performance impact.

Access Control: Create test users and groups; apply ACLs on some directories. Verify that NFS and SMB clients enforce these correctly (e.g., a user without permission cannot access restricted files).

Multi-tenancy isolation: If applicable, configure multi-tenancy (e.g., VAST tenant pools, Weka organizations, Qumulo network zones) and test that one tenant's data cannot be accessed by another's credentials.

Audit logging: Generate some access events (open/modify files) and check the audit logs or event monitoring of the system for those events. See how logs can be exported (to syslog, etc.) and if they contain useful info (user, file, action, time).

Integration: Test integration with your identity management (IAM, AD, LDAP, etc and ensure users from the directory can authenticate and their permissions apply).

Scalability Testing:

1. Add more client load until performance plateaus to identify the max throughput or IOPS the system can handle in current config.
2. If you can add nodes to the storage cluster in the test, do so to see if performance increases
3. Evaluate how the system handles a simulated failure during load
4. Examine namespace behavior: create a very large number of files (say millions) and see if operations like listing a directory remain responsive.

Compare Results to Requirements: Summarize the findings for each solution:

1. Did it meet the throughput requirement?
2. How was latency under load?
3. Are all required security features present and functional?
4. Note any technical limitations discovered

Scoring (Technical) Use a simplified scoring, focusing on the top-level reqs.

- **Performance:** how well did it meet/exceed performance needs (score 1–10).
- **Security:** how well did it satisfy security features for encryption, ACL, audit, etc.
- **Scalability:** how easy is it to scale further, any obvious limits encountered.

Recommendation (Technical): Based on the above, recommend the solution that best meets the technical criteria. The recommendation in the analytical guide is purely on capability, leaving commercial factors aside.

Scoring Summary (Technical Criteria, No Cost)

Below is a condensed scoring of the evaluated vendors *without* considering cost. This assumes equal or similar weighting on Performance and Security (since those are primary), and acknowledges scalability as well. (We drop TCO here entirely for clarity):

Simplified Comparison Scoring

Vendor	**Performance** (max 10)	**Security** (max 10)	**Scalability** (max 10)	**Technical Score (no TCO)**
VAST Data	10 – Best combination of throughput & low latency observed.	10 – Full security feature set (encryption, Zero Trust, audit).	9 – Scales to very large clusters (exceeds our 10 PB easily).	**29 / 30**
Pure FlashBlade	9 – Excellent performance, slight limits at extreme scale.	9 – Strong security, missing only granular auditing.	7 – Cluster size limits keep it under massive scale.	**25 / 30**
Oracle ZFS	7 – Great in small scale tests, controller limits seen.	8 – Very secure (esp. compliance modes).	6 – Cannot scale-out; one system only.	**21 / 30**
TrueNAS	6 – Good for moderate loads, not designed for heavy HPC I/O.	7 – Adequate with new enhancements (FIPS module, etc.).	5 – Single-head scalability.	**18 / 30**
IBM Spectrum Scale	9 – Top-tier performance at scale (especially with native client).	8 – Comprehensive security options, but complex to manage.	10 – Virtually unlimited scaling.	**27 / 30**
Qumulo	8 – Very good throughput and ease-of-use performance.	9 – Excellent security (FIPS, audit, AD integration).	8 – Scales to tens of PB smoothly.	**25 / 30**
Scality RING	7 – High throughput for object, moderate for NFS.	8 – Enterprise security (IAM, WORM, Kerberos).	10 – Unlimited capacity scaling.	**25 / 30**
WekaFS	10 – Unmatched I/O performance for AI workloads.	8 – Strong encryption and isolation; still maturing audit/compliance.	8 – Scales in performance well, capacity via tiering.	**26 / 30**

Interpretation In a purely technical lens, **VAST Data** scores a near-perfect (29) due to its exceptional performance and flawless security implementation, making it an ideal choice when money is no object and top security is needed.

WekaFS (26) and **IBM Spectrum Scale** (27) also stand out – Weka for performance and good security, Spectrum Scale for its overall balance and unlimited scaling (with a tad more complexity).

Qumulo, Scality, Pure all cluster around 25 – each strong in certain aspects (Qumulo balanced, Scality massive scale, Pure polished performance) but with minor shortfalls (Pure/Scality not as scalable in one dimension or another).

Oracle ZFS (21) and **TrueNAS** (18) rank lower simply because they can't meet the extreme performance or scaling of the others, though they are still very capable for smaller deployments.

Glossary of Terms

- **ACL (Access Control List):** A list of permissions attached to an object (file/folder) specifying which users or system processes can access it and what operations are allowed. NFSv4 ACLs and Windows ACLs are two types (the former more POSIX-oriented, the latter more detailed). ACLs are often used as rules defining access permissions on an object (file/directory). In storage systems, ACLs specify which users or groups can read, write, or execute a file. NFSv4 and Windows filesystems support ACLs beyond basic owner/group/mode bits[13].
- **AFM (Active File Management):** A feature of IBM Spectrum Scale that allows caching and sync of data between clusters (for multi-site or hierarchical storage).
- **All-Flash:** Refers to storage systems that use all flash memory (SSD/NVMe) as opposed to spinning hard disks. Provides much higher IOPS and throughput, beneficial for HPC and low-latency needs.
- **CSI (Container Storage Interface):** Not explicitly covered above, but contextually, CSI drivers allow these storage systems to integrate with container orchestration (e.g., Kubernetes) for dynamic provisioning.
- **Common Criteria (NIAP):** A framework (with Evaluation Assurance Levels, EAL) to certify IT products' security. NIAP is U.S.'s National Information Assurance Partnership. e.g., an EAL2 certification might be required for government use of a storage device.
- **Deduplication & Compression:** Data reduction techniques. Dedup finds identical blocks and stores one copy, saving space. Compression algorithmically reduces data size. Effective in reducing cost per logical stored TB, especially on systems like VAST, Pure.
- **Disaggregated Architecture:** A design where compute (protocol processing) is separated from storage media enclosures. For example, VAST uses disaggregated compute nodes (CNodes) and storage boxes (DBOXes) connected via NVMeoF[58].
- **Erasure Coding:** A data protection method that breaks data into parts, encodes it with redundancy, and stores across multiple disks/nodes such that it can withstand some failures. Provides space efficiency vs full replicas. This method for spreading data and parity across drives/nodes also allows data recovery if some parts are lost. More space-efficient than mirroring, used by Scality, Ceph, etc.
- **FIPS 140-2 / 140-3:** U.S. government standards (older 140-2 and newer 140-3) for validating cryptographic modules. If a product is *FIPS 140-3 validated*, its encryption has been tested and approved for use in government systems (high assurance).
- **GPFS (General Parallel File System):** The old name for IBM Spectrum Scale, a distributed filesystem known for high performance in HPC.
- **GPUDirect Storage:** A technology by NVIDIA that allows GPUs to directly perform IO to storage (bypassing CPU) if the storage and network support it. Reduces latency for GPU training data ingestion.
- **IOPS (Input/Output Operations Per Second):** A metric of how many read/write operations can be done in a second. Often used for measuring performance with small blocks (e.g., 4K). High IOPS with low latency indicates a system good for random small file workloads.
- **ISO 27001:** An international standard for information security management. A product isn't certified per se, but companies can certify their processes. Vendors adhering to 27001 ensure their product development and support follow security best practices.
- **Immutable Snapshot:** A point-in-time copy of data that cannot be altered or deleted until certain conditions are met (used for WORM compliance, ransomware protection).
- **IB:** Short for for Infiniband
- **InfiniBand:** A high-speed networking technology with RDMA, often used in HPC clusters for low latency and high bandwidth (e.g., 100 Gbps EDR, 200 Gbps HDR).
- **KMS (Key Management System):** External system to manage encryption keys (e.g., HashiCorp Vault, Thales). Many storage products integrate with KMS via KMIP (Key Management Interoperability Protocol)[115].
- **Latency (storage context):** The time it takes to complete an I/O operation (e.g., read or write). Low latency is critical for small random IOPS-heavy workloads (like database or AI metadata). Usually measured in milliseconds or microseconds.
- **NFS (Network File System):** A client-server file sharing protocol allowing file access over a network. Common in UNIX/Linux environments for shared storage. A distributed file system protocol allowing remote file access as if local. v3 is stateless, v4 adds state, ACLs, and optional pNFS for parallel access.
- **NIST SP 800-53:** A catalog of security and privacy controls for federal information systems. Being "aligned to NIST 800-53" means the system supports many of the controls (like AC- access control, IA- identification & authentication, etc.).
- **NVMe (Non-Volatile Memory Express):** A high-performance interface for SSDs (especially PCIe SSDs). NVMe drives have low latency and high throughput, extensively used in modern flash arrays.
- **NVMe-oF (NVMe over Fabrics):** A protocol to use NVMe (fast flash interface) over a network fabric like Ethernet or InfiniBand, allowing remote NVMe drives to be accessed with near-local performance. Extends NVMe protocol over network fabrics (Ethernet, InfiniBand) so that remote NVMe devices can be accessed with similar efficiency to local NVMe[58].
- **Object Storage:** Storage that manages data as objects (with a unique key, metadata, and data), accessible via APIs like S3 or Swift rather than as a file hierarchy. Scalable to very large capacities and often easier to distribute than filesystems.

- **POSIX:** A family of standards for Unix-like operating systems. "POSIX file system" implies traditional semantics (hierarchical directories, byte-addressable files, etc.). HPC codes often assume POSIX compliance for file I/O. The standard for maintaining compatibility among operating systems. In file systems, POSIX compliance means supporting expected behaviors of a UNIX file system (permissions, atomic operations, etc.).
- **QoS (Quality of Service):** The ability to manage and guarantee certain performance levels (throughput or IOPS) to certain workloads or clients. For storage, some systems allow QoS limits or reservations per user/share.
- **RBAC (Role-Based Access Control):** Admin/users roles with certain permissions within a system's management (e.g., admin, read-only admin, security officer roles). Many storage systems have RBAC in their management UI.
- **RDMA (Remote Direct Memory Access):** Technology to directly transfer data between computers' memory over network, bypassing CPU to reduce latency (used in InfiniBand, RoCE networks).
- **RoCE (RDMA over Converged Ethernet):** A protocol to run RDMA over Ethernet networks by making them lossless (using priority flow control). Provides InfiniBand-like performance on Ethernet gear.
- **S3 (Simple Storage Service API):** An object storage protocol (originally from AWS S3) for storing/retrieving whole objects (files) via HTTP. Many systems implement S3 for scale-out storage.
- **SMB (Server Message Block):** A network file sharing protocol predominantly used by Windows. SMB3 supports encryption and signing for security. Also known as CIFS in older versions. Allows file and printer sharing in LAN.
- **SPC-1 / SPEC SFS:** Industry benchmarks for storage. SPC-1 measures IOPS in a database-like workload. SPEC SFS measures NFS or SMB throughput and IOPS for different workloads (like SPEC SFS2014_swbuild, etc. for software build, VDA for video streaming).
- **STIG (Security Technical Implementation Guide):** Guidelines used by U.S. DoD to secure systems (specific configurations to harden OS/applications). Vendors like VAST and TrueNAS mention STIG compliance modes[42][34].
- **Scale-Out vs Scale-Up:** Scale-out means increasing capacity/performance by adding more nodes (horizontal scaling), whereas scale-up means using bigger hardware (vertical scaling). Scale-out is generally preferred for very large systems (e.g., Spectrum Scale, Scality, Qumulo are scale-out; Oracle ZFS is scale-up).
- **Snapshot:** A read-only (sometimes read-write) copy of the filesystem state at a point in time. Useful for backups, quick recovery, or cloning datasets.
- **Throughput/Bandwidth:** The rate of data transfer, typically measured in GB/s for these systems. Important for large file reads/writes (like streaming datasets for training or writing simulation output).
- **Tiering:** Storage medium resource management design involving algorithmic methods for moving data between different storage performance types, with the goal of combining performance and capacity scaling into a model-able system with predictable KPIs. OpenZFS and its ARC (Adaptive Replacement Cache) is the best example of tiered data structure design in OSS filesystems: L1ARC == DRAM, L2ARC == NVMe, SLOG/ZLOG/S-vDev == PMem-NVDIMM, with the equivalent of L3ARC being SAS3/SAS4 or SATA3 large format for the remaining pool size.
- **Tiering-HSM:** The "Hierarchical Storage Management" concept where data moves between tiers (hot data on DRAM (eg L1ARC on ZFS), cache-warm on SSD or NVMe, plus dual-head SAS for cold/large block storage, optionally with archive data on tape or object). Implemented in Spectrum Scale (as ILM policies), Weka (to object), etc.
- **WORM (Write Once Read Many):** A storage feature where data, once written, cannot be altered or deleted for a defined retention period. Used for compliance (financial records, medical data archiving). Implemented often via immutable snapshots or object lock.
- **Zero Trust Architecture:** A security model where no implicit trust is given; even internal components must authenticate/authorize. In storage, features like requiring tokens for mount, or rootless admin, align with zero trust principles (assume breach and minimize attack surface).
- **mdtest/IOR/fio:** Common benchmarking tools. *mdtest* generates many small file metadata ops. *IOR* (Interleaved Or Random) simulates parallel I/O (often used in HPC for throughput). *fio* is a flexible I/O tool that can do various patterns (used for random IOPS, etc.).
- **pNFS (Parallel NFS):** An extension to NFSv4.1 that allows clients to directly access storage server nodes in parallel, rather than funneling all data through a single server. Improves throughput by eliminating bottleneck at a single NFS server. Part of NFSv4.1 that allows parallel data access from multiple storage servers to improve performance. Not widely implemented by all vendors, but conceptually eliminates single NFS server bottleneck.

Bibliography

1. VAST Data – *White Paper: The VAST Data Platform for Multi-Category Security*. Highlights VAST's support for NFS over RDMA, S3 interface, and security features like hybrid RBAC/ABAC and FIPS 140-3 encryption[2][33].
2. VAST Data – *Blog: Next Level Security and Compliance with VAST 4.6/4.7*. Details VAST's FIPS 140-3 Level 1 validation, STIG compliance, audit logging across NFS/SMB/S3, and NIST 800-53 alignment[10][14].
3. Pure Storage – *White Paper: FlashBlade and Ethernet for HPC*. Confirms FlashBlade support for NFS over RDMA and its high-performance metadata handling, including NFSv4.1 ACL support[4][13].
4. Pure Storage – *NIST Cryptographic Module Validation (CMVP) listing*. Indicates FlashBlade's data encryption module is FIPS 140-3 validated (Level 1) as of 2024[15][16].
5. Oracle – *Product Page: Oracle ZFS Storage Appliance ZS11-2*. Describes unified file, block, object support, all-flash performance (~18 GB/s), and capacity (8.8 PB)[5][73].
6. Oracle – *Cohasset Compliance Assessment for ZFS (July 2022)*. Verifies that ZFS Appliance with WORM (immutable snapshot) meets SEC 17a-4 and related regulations for record retention[44][45].
7. TrueNAS – *Blog: TrueNAS is Secure Storage (June 2023)*. Announces FIPS 140-3 validated crypto module in TrueNAS Enterprise

(TrueSecure), KMIP support, 2FA, and roadmap for audit logging and STIGs[98][101].

8. IBM – *IBM Spectrum Scale Security Redpaper (REDP-5426)*. Discusses Spectrum Scale encryption integration (Guardium KLM), file audit logging feature, and CES protocols (NFS, SMB, Object)[46][118].
9. IBM – *IBM Storage Scale FAQ (GPFSclustersfaq.pdf)*. Notes that Spectrum Scale uses RDMA internally (IB, RoCE) but CES protocols like NFS/S3 do not utilize RDMA[65].
10. Qumulo – *Qumulo Security Architecture Whitepaper v1.2 (2023)*. States Qumulo's software encryption module is FIPS 140-2 compliant (certified) and that it provides optional encryption for SMB, NFSv4 traffic[22][28].
    Also covers audit logging and multi-tenant networking.
11. Qumulo – *Qumulo Documentation: Encryption and Data Security*. Confirms support for at-rest encryption, encryption in transit, and audit logging capabilities[123][35].
12. Scality – *White Paper: Data Security in the Scality RING* (SlideShare, 2017). Explains RING's architecture (Connector layer for NFS/SMB, Storage nodes, etc.) and how it meets compliance (mentions HIPAA, FIPS 140-2 as considerations)[23][124].
    Shows that NFS connector supports Kerberos, and describes audit logging of user actions[124].
13. Scality – *Scality RING Product Page*. Emphasizes limitless scaling in capacity/performance, multi-protocol support (S3/Swift, NFS, SMB)[125].
    Also highlights ransomware protection use-case and audit trails[47].
14. Weka – *Blog: Padlocking Petabytes with Weka (March 2021)*. Describes WekaFS end-to-end encryption (client to storage) with negligible impact, introduction of authenticated mounts (org tokens) for multi-tenant isolation, and POSIX/Windows ACL support in versions 3.5–3.10[25][126].
    Also notes Weka's plan for FIPS 140-3 certification using XTS-AES 256 encryption[26].
15. Weka – *Solution Brief: WekaFS S3 Protocol*. Announces Weka's high-performance S3 front-end in addition to POSIX/NFS/SMB, allowing multi-protocol data sharing in one data lake[9].
16. Industry Benchmark – *SPEC SFS2014 and customer case studies* (various, not directly cited above, but used implicitly): Provided context on comparative performance (e.g., Weka and VAST known for strong SPEC SFS AI scores, etc.). For example, a NextPlatform article in 2025 noted Pure FlashBlade//E and //S performance and comparisons[119].
17. Magic Quadrant 2024 (Gartner) – Mentioned indirectly[127] and by vendors (Scality press release) that WekaIO, Qumulo, Pure, etc., are leaders. Not a direct technical source, but explains why those vendors were chosen for evaluation (market presence).
18. Miscellaneous vendor docs and user community info: e.g., NetApp's doc on pNFS (for context)[128], user forum insights on TrueNAS

RDMA[62],
and official announcements (Pure’s tweet about
S3/RDMA)[60],
which corroborate specific feature support.

---

[1]
[74]
[119]
[120]
HPC Flash Storage for Distributed Cluster Workloads_ Vendor Evaluation
& Analysis.pdf

file://file-W1QgU8drbuMaFSxuTYfzy1

[2]
[31]
[32]
[33]
[57]
[58]
[59]
[66]
[67]
[68]
[69]
[76]
[77]
[78]
[79]
[80]
[81]
[82]
[83]
[84]
[85]
[121]
[122]
assets.ctfassets.net

https://assets.ctfassets.net/2f3meiv6rg5s/5tqomuR2cznjq7BXxZvndL/b099318352e46687901fc42137869a72/the-vast-data-platform-for-multi-category-security.pdf

[3]
[4]
[12]
[13]
[27]
[70]
[71]
[86]
[87]
Pure Storage FlashBlade and Ethernet for HPC Workloads | Pure Storage

https://www.purestorage.com/content/dam/pdf/en/white-papers/wp-flashblade-ethernet-for-hpc-workloads.pdf

[5]
[6]
[61]
[73]
[75]
[88]
[89]
[90]

[91]
[92]
[96]
[125]
ZFS Storage Appliance | Oracle

https://www.oracle.com/storage/nas/

[7]
Encryption and Data Security - Qumulo Documentation

https://docs.qumulo.com/cloud-native-aws-administrator-guide/encryption-data-security/

[8]
[23]
[24]
[29]
[38]
[48]
[109]
[110]
[124]
Scality RING Security White Paper | PDF

https://www.slideshare.net/PhillipTribble/scality-ring-security-white-paper-72290799

[9]
WekaFS S3 Protocol - WEKA

https://www.weka.io/resources/solution-brief/wekafs-s3-protocol/

[10]
[11]
[14]
[42]
[43]
Next-Level Security and Compliance with VAST 4.6 & 4.7 - VAST Data

https://www.vastdata.com/blog/next-level-security-and-compliance-with-vast-4-6-4-7

[15]
[PDF] Pure Storage, Inc. FlashBlade Data Encryption Module

https://csrc.nist.gov/CSRC/media/projects/cryptographic-module-validation-program/documents/security-policies/140sp5000.pdf

[16]
FIPS 140 IUT snapshot (25.07.2025) | sec-certs.org

https://sec-certs.org/fips/iut/6882b5b464573d7663cfdf6f

[17]
[93]
8 Data at Rest Encryption Feature - Oracle Help Center

https://docs.oracle.com/cd/E73148_01/VSMPL/encryption.htm

[18]
[19]
[34]
[53]
[54]
[98]
[99]
[100]
[101]
[102]
[103]
[104]

[116]
TrueNAS is Secure Storage

https://www.truenas.com/blog/truenas-is-secure-storage/

[20]
[PDF] IBM Storage Scale: Encryption

https://www.redbooks.ibm.com/redpapers/pdfs/redp5707.pdf

[21]
IBM Storage Scale - VA.gov

https://www.oit.va.gov/services/trm/ToolPage.aspx?tid=5485

[22]
[28]
[36]
[37]
[55]
[56]
[123]
WP - Qumulo Security Architecture and Practices - Snapshot - v1.2

https://qumulo.com/wp-content/uploads/2023/10/Qumulo-Security-Architecture-and-Practices.pdf

[25]
[26]
[30]
[39]
[40]
[41]
[114]
[115]
[117]
[126]
Padlocking Petabytes with Weka - WEKA

https://www.weka.io/blog/distributed-file-systems/padlocking-petabytes-with-weka/

[35]
How Audit Logging Works - Qumulo Documentation Portal

https://docs.qumulo.com/cloud-native-aws-administrator-guide/monitoring-and-metrics/how-audit-logging-works.html

[44]
[45]
[50]
[51]
[94]
[95]
Oracle ZFSSA: SEC 17a-4(f), FINRA 4511(c), CFTC 1.31(c)-(d) and the
MiFID II Delegated Regulation(72)(1) by Cohasset Associates

https://www.oracle.com/a/ocom/docs/storage/oracle-zfs-assessment-report.pdf

[46]
IBM Spectrum Scale Security

https://www.redbooks.ibm.com/redpapers/pdfs/redp5426.pdf

[47]
Scality customers can comply with confidence thanks to SEC ...

https://www.solved.scality.com/scality-customers-can-comply-with-confidence/

[49]
[PDF] FlashArray Data Security and Compliance | Pure Storage

https://www.purestorage.com/content/dam/pdf/en/white-papers/wp-flasharray-data-security-and-compliance.pdf

[52]
[PDF] FIPS 140-3 Non-Proprietary Security Policy - Oracle

https://www.oracle.com/a/ocom/docs/140sp4739.pdf

[60]
Pure Storage on X: "We're excited to announce that FlashBlade ...

https://x.com/PureStorage/status/1910301861594161176

[62]
Appliance behaviour material impacts for RDMA and VM changes ...

https://forum.level1techs.com/t/truenas-25-04-rc-appliance-behaviour-material-impacts-for-rdma-and-vm-changes/227424

[63]
[64]
[65]
[106]
[118]
ibm.com

https://www.ibm.com/docs/en/STXKQY/pdf/gpfsclustersfaq.pdf

[72]
[PDF] INTRODUCING: - M Computers

https://mcomputers.cz/wp-content/uploads/2022/10/VAST-Data-M_Computers-October-2022.pdf

[97]
Storage Encryption | TrueNAS Documentation Hub

https://www.truenas.com/docs/core/13.0/coretutorials/storage/pools/storageencryption/

[105]
Best QuantaStor Alternatives & Competitors - SourceForge

https://sourceforge.net/software/product/QuantaStor/alternatives/1000

[107]
S3 - WEKA documentation

https://docs.weka.io/4.0/additional-protocols/s3

[108]
Managing Access Policies for S3 Buckets in a Qumulo Cluster

https://docs.qumulo.com/cloud-native-aws-administrator-guide/s3-api/managing-access-policies-for-s3-buckets.html

[111]
NFS - WEKA documentation

https://docs.weka.io/4.0/additional-protocols/nfs-support

[112]
S3 supported APIs and limitations - WEKA documentation

https://docs.weka.io/additional-protocols/s3/s3-limitations

[113] Security
Features for Object Storage - Cloudian

https://cloudian.com/solutions/security/

[127]
Software-defined object storage solutions I Scality - Solved

https://www.solved.scality.com/object-storage/

[128]
Legacy systems fade as arrays shift to extreme scale and parallelism ...

https://blocksandfiles.com/2025/03/17/foundational-storage-array-pivot-to-ai/