Parallel File System Size Recommendation for GPU Clusters

1. Purpose

This document provides sizing guidance for the parallel file system (PFS) supporting GPU-accelerated AI/ML workloads. The recommendations are based on vendor best practices, performance reference architectures, and observed usage patterns from comparable deployments. The sizing serves as a baseline; actual requirements must be adjusted based on business needs, workload profiles, and growth projections.

2. Background

A GPU cluster’s performance is heavily dependent on the storage system’s ability to deliver sustained throughput and low latency at scale. In AI/ML environments, the PFS (e.g., DDN EXAScaler, WEKA FS, VAST Data, Lustre) must: - Deliver high bandwidth for model training and large dataset ingestion. - Provide low-latency small file access for metadata-heavy operations. - Scale linearly as GPU count and dataset sizes grow. The storage sizing is typically driven by per-GPU flash capacity and per-SuperPOD throughput guidelines, as outlined by vendors such as NVIDIA, DDN, WEKA, and VAST.

3. Methodology

3.1 Sizing Drivers

1. Per-GPU Flash Capacity - NVIDIA DGX SuperPOD RA and customer deployments indicate a typical range of: - 1–1.5 TB flash per GPU for balanced cost/performance. - 3–4 TB per GPU in high-retention or multi-tenant environments. - Factors influencing capacity: - Dataset size and retention policy. - Number of concurrent experiments. - Checkpoint frequency. - Data replication or erasure coding overhead. 2. Throughput Requirements - NVIDIA DGX SuperPOD recommends: - 125 GB/s per SuperPOD Unit (SU) (8 nodes) for sustained training. - 15–20 GB/s per DGX node as a planning baseline. - Storage must maintain line-rate performance during concurrent reads/writes. 3. Performance Headroom - Add 20–30% capacity headroom for metadata growth, burst training phases, and temporary datasets. 4. Redundancy and Resiliency - Overprovision capacity to account for RAID/erasure coding overhead (10–30%) and maintenance operations.

4. Vendor Reference Guidance

Vendor / Reference	Capacity Rule of Thumb	Throughput Guideline	Notes
NVIDIA DGX RA	1–1.5 TB flash per GPU	125 GB/s per SU	8 DGX per SU
DDN EXAScaler	1–2 TB per GPU typical	20 GB/s per DGX node	Uses Lustre parallel FS
WEKA FS	1–2 TB per GPU typical	150–200 GB/s per SU	All-flash tiering, object backend optional
VAST Data	2–3 TB per GPU	125 GB/s per SU	Disaggregated, scale-out flash

5. Sizing Example

Scenario: - Cluster: 4 SuperPOD Units (SU) - GPUs per SU: 64 (8 DGX nodes × 8 GPUs each) - Total GPUs: 256 - Per-GPU Flash Target: 1.5 TB Calculation: - Raw Flash Capacity = 256 GPUs × 1.5 TB = 384 TB - Add 25% Headroom = 480 TB usable target - Account for Erasure Coding (20%) = ~600 TB raw provisioned Throughput Requirement: - 125 GB/s per SU × 4 SU = 500 GB/s aggregate sustained throughput

6. Final Recommendation

Capacity: - Provision ~600 TB raw flash to achieve 480 TB usable for initial phase. - Plan for linear expansion to 1 PB raw by Year 2 to accommodate dataset growth and additional GPUs. Performance: - Ensure 500 GB/s sustained throughput capability, with sub-millisecond latency for metadata operations. - Maintain balanced performance across read-heavy, write-heavy, and mixed workloads. Growth Plan: - Year 1: Size for baseline workloads + 30% buffer. - Year 2–3: Expand in line with GPU additions or dataset expansion.

7. Key Assumptions

- Workloads are AI/ML training and inference with mixed I/O patterns. - Checkpoint data is stored on PFS; archival data migrates to object storage. - PFS serves multiple teams but is not a multi-tenant regulated environment (which would require higher per-GPU capacity).

8. References

1. NVIDIA DGX SuperPOD Reference Architecture – Storage Architecture Guidelines
2. DDN EXAScaler Deployment Best Practices
3. WEKA FS Sizing Guide for AI/ML Clusters
4. VAST Data AI/ML Deployment Considerations