GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

HPC Storage at GWDG

Experiences from the past and planning for the future

Sebastian Krey

25.09.2024

Outline

Storage Systems: Current

- HOME/SW: 350 TiB DDN Gridscaler, EoL 08/24
- WORK MDC: DDN ExaScaler 5 EoL 08/24
  - Metadata SFA7700X
  - 8 PiB HDD 2x ES14KX
  - 113 TiB NVME 2x SFA200NV
- WORK RZGÖ: DDN ExaScaler 6 113 TiB NVME 2x ES400NVX
- HOME/SW/WORK KISSKI: VAST Data 500TiB NVME (1x dBox, 2x cBox)
- WORK SCC: 2.2 PiB BeeGFS based on DDN SFA7990 block storage
- HOME SCC: 3 PiB Quantum StorNext
- HSM/Tape: Quantum StorNext HSM 3 PiB (EoL 01/25)

Current storage concept

- Different user groups have different storage systems available
- The same path (e.g. `/scratch`) can point to filesystems with different characteristics.
- Not all storage systems are available on all nodes
- Different concepts for data sharing depending on source of project/user (compute projects, functional accounts, etc.)
- Unified operation requires same storage access for all nodes and currently not possible accross all systems
- Users of Tier 3 system have their campus home as home directory

New unified storage concept for NHR/SCC/KISSKI

- Replace HDD based WORK storage with central Ceph instance
- Compute island specific high performance storage, all flash
  (Lustre, VAST or BeeGFS, DAOS maybe a candidate in the future)
- Unify HOME/SW to central HPC home storage
- HPC S3 object storage for “Cloud” workloads and easy data ingest/export with central S3 storage of infrastructure group and external parties
- Access to campus home directory (StorNext) only via data mover nodes
- Semantic storage: Assignment of storage backend based on project requirements, transparent access via symlinks.
- Directory quotas, whenever possible

Storage assignment

- Based on project application space and filesystem type will be assigned
- Every user gets home directories for their project specific user accounts
- Every project gets their volume storage in the central coldstorage
- Every project gets archive storage based on requirements
- In RZGÖ assingment of high performance storage based on I/O requirements (Lustre or VAST depending on read/write mix)
- Open question: Management of campaign storage
  - Admin assignment or self management by user/project $\rightarrow$ Workspaces

Problems with current storage

Lustre: long failover times and crashes happen quite often, enterprise support usually way behind open source version regarding kernel support

BeeGFS: lacks some features like project/directory quota, performance scaling in larger NVME setups

GPFS: expensive, strict kernel version requirements, no directory quota, metadata handling on client can be advantage and disadvantage, limited fabric support

StorNext: architecture outdated (focus on SAN, single MDS), slow bugfixing and updates for kernel support, current licensing model expensive, missing a lot of modern features

Strengths of the different storage technologie

VAST:	Extreme high availability, NFS everywhere useable, high read speed, consistent low latencies
Lustre:	Extreme high performance possible, user configurable striping
BeeGFS:	Very easy to setup and manage, good performance
GPFS:	Can a lot of stuff, good performance possible
Ceph:	Capacity scaling, very low EUR/TB, properly setup: good performance

Storage Systems: New Homestorage (in procurement)

- Unified home storage for all user groups
- Expansion of existing VAST storage
- 600 TiB of all flash storage
- Mounted via NFS on all compute nodes
- Will also provide the central software installation
- Strict volume quota, relaxed inode quota
- Daily snapshots and offsite backup

Storage Systems: New High Performance storage (in procurement)

- Expansion of WORK RZGÖ (Lustre) by replacing 4TB SSDs with 15TB SSDs
- New (likely) Lustre based filesystem for WORK MDC (approx 1-1.5PiB)
- Usage limited to specific compute island to ensure high performance
- Strict volume and inode quota
- All flash filesystems to allow best performance in all workload types
- Smaller HPC hosting: BeeGFS for easy setup and maintenance

Storage Systems: New Coldstorage (in setup phase)

Hardware:

- 53 Servers, 21 PB HDD, 3.5 PB NVME
- HDD Cluster with 45 Servers:
  - 24x 20TB HDD, 4x 7.68 NVME
  - 2x24 Core Sapphire Rapids CPUs, 512 GB memory
  - 2x25G Ethernet
- NVME Cluster with 8 Servers
  - 20x 15.36TB NVME
  - 2x32 Core Milan CPUs, 512GB memory
  - 100G Ethernet
- HDD cluster capacity optimized $\rightarrow$ Erasure Coding
- NVME cluster performance optimized $\rightarrow$ Replication
- Installation support from “Clyso”

Ceph for HPC?

Common opinion:

- Are you insane?
- Ceph is slow, complex, unreliable,. . .
- Only TCP connections

On closer look:

- Ceph is reliable standard in cloud environments
- Some institutes use it successfully in HPC (e.g. CERN, IZUM)
- Ceph allows complete hardware vendor independence
- Hardware migrations in live operation, without user interaction
- Recent performance improvements show respectable performance (work from Clyso and Croit)
- With enough CPU cores and memory 75-80% network saturation
- Enough MDS containers achieves very good metadata performance scaling

Storage for GPU cluster/Deep Learning

Myths:

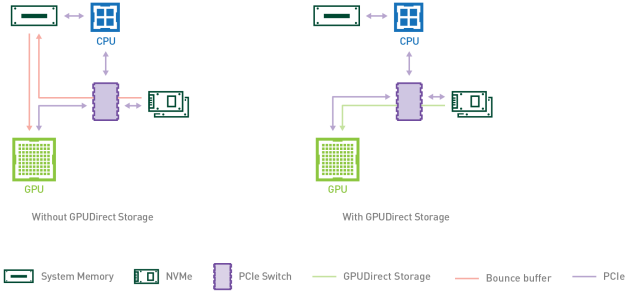
- Insanely high metadata performance necessary
- IOPs explosion
- Multi GB/s per GPU necessary
- Only GPUDirect Storage enabled storage can satisfy GPU workloads

Reality:

- Metadata and IOPs explosion usually result of bad IO design
- First training epoch requires high performance, then caching possible
- High performance for first GPU node but scaling for cluster size installations not heavier than for CPU systems

What is GPUDirect Storage

Nvidia GPUDirect Storage (GDS) or Nvidia Magnum IO provides a direct DMA path between GPU and PCIe attached storage via the `cuFile` API in a Nvidia ConnectX-4+ based fabric.



Source: https://developer.nvidia.com/blog/gpudirect-storage/

Examples for useable storage: local NVME drives, Lustre, BeeGFS, GPFS, WekaFS, VASTData, NetApp ONTAP, RDMA enabled NFS

Results 4

Bypassing GDS for blocksizes <1MB on Lustre with 2 nodes:

Lustre 2 nodes	cuFile (POSIX)	cuFile (GDS)
4k rand read IOPS	1.933.847	2.215.756
512k rand read MiB/s	57.396	55.650
GPU utilization %	>80	<5

Throughput very similar but large difference in GPU utilization during IO.

DGX Superpod Design Guide

Table 6. Guidelines for storage performance

Performance Characteristic	Good (GBps)	Better (GBps)	Best (GBps)
Single-node read	4	8	40
Single-node write	2	4	20
Single SU aggregate system read	15	40	125
Single SU aggregate system write	7	20	62
4 SU aggregate system read	60	160	500
4 SU aggregate system write	30	80	250

Source https://docs.nvidia.com/https:/docs.nvidia.com/dgx-superpod-reference-architecture-dgx-h100.pdf

DGX Superpod Design Guide

Table 5. Storage performance requirements

Performance Level	Work Description	Dataset Size
Good	Natural Language Processing (NLP)	Datasets generally fit within local cache
Better	Image processing with compressed images (ex: ImageNet)	Many to most datasets can fit within the local system's cache
Best	Training with 1080p, 4K, or uncompressed images, offline inference, ETL,	Datasets are too large to fit into cache, massive first epoch I/O requirements, workflows that only read the dataset once

Source https://docs.nvidia.com/https:/docs.nvidia.com/dgx-superpod-reference-architecture-dgx-h100.pdf

Summary

GWDG perspective:

- Target of larger vendor independance will be achieved
- Number of technologies will be reduced
- Unified storage platform for different user requirements

Storage general:

- Storage for GPU based systems is not as hard as thought
- Not all features of vendor marketing are really needed
- Ceph better suitable for HPC requirements than imagined
- A single storage system for everything is difficult