GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

The (new) storage systems at GWDG

Sebastian Krey

07.08.2024

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 (compute projects, functional accounts, etc.)
- Unified operation requires same storage access for all nodes and currently not possible accross all systems

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 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

Storage Systems: New Coldstorage (in setup phase)

Hardware:

- 53 Servers, 23 PB HDD, 3.5 PB NVME
- HDD Cluster with 45 Servers:
  - 24x 22TB 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”

Storage Systems: New Homestorage (in procurement)

- Unified home storage for all user groups
- Likely expansion of existing VAST storage
- 400 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

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:
  - Fixed high performance storage assignment for every project
  - Workspace solution (self allocation of high performance storage for a limited time)

Storage migration required

- As soon as the new storage systems are production ready we will announce the upcoming data migration
- Open question:
  - Individual user based data migration
  - Admin managed data migration
  - Duration of migration period (3 months?)
- After the end of the migration shutdown of old storage systems, no data recovery possible anymore

Summary

- Old storage systems will be replaced by larger and faster storage in the next months
- Data migration will be necessary
- Data access for some user groups will change (esp. SCC users)
- Unified operational concept will allow easier migration from Tier 3 (SCC) to Tier 2 (NHR) usage for university users
- Easier maintenance and documentation will allow a better user experience, performance and availability