GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Data management in science

Challenges through the demands of HPC and AI

Sebastian Krey, Julian Kunkel, Christian Boehme

02.06.2022

Outline

Introduction GWDG

- Data and IT service center for **University of Göttingen** and **Max Planck Society** (MPG) since 1970
- Since 2021 working group “Computing” for HPC operation.
- Operating site for the “North-German Supercomputing Alliance” (HLRN) since 2018, since 2021 in transition to NHR
- HPC operating site for the “German Aerospace Center” (DLR) since 2022

HPC systems at GWDG

- Tier 3: **Scientific Compute Cluster** (SCC)
- Tier 2: **HLRN-IV “Emmy”**
- **“CARO”** for **German Aerospace Center** (DLR)

Scientific Compute Cluster (SCC)

Overview

- local HPC system of GWDG (Tier 3)
- accessible to users from Uni Göttingen and MPG
- tight integration with GWDG infrastructure (User management, Home- and S3 Objectstorage, etc.)
- several islands (CPU and GPU partitions with different storage and interconnect requirements) distributed across three data centers

- specs summary:
  - 300 nodes
  - 270 GPUs, e.g.: GTX1080, RTX5000, Tesla V100
  - Homestorage: central StorNext installation
  - Workstorage: BeeGFS filesystems in each compute island
  - Objectstorage: central Ceph clusters
  - Archive: central StorNext HSM solution with Quantum iScalar tape libraries

HLRN-IV “Emmy”

Overview Tier 2 system

- named in honor of *Emmy Noether*, German mathematician
- TOP 500: #47 in 2020-11 (5.95 PFLOP/s) (with SCC expansion)
- phase 1 compute nodes
  - 2x Intel Xeon Gold 6148 (SKL), 40 cores per node, 480 GB SSD
  - 432x 192 GB, 16x 768 GB
  - 3 GPU nodes: 2x Intel Xeon Gold 6148, 768 GB, 4x Tesla V100 32 GB
- phase 2 compute nodes
  - 2x Intel Xeon Platinum 9242 (CLX-AP), 96 cores per node
  - 3004x 384 GB, 16x 768 GB, 2x 1536 GB
- Upcoming expansion with 36 GPU nodes (2xEpyc 32 cores, 4xA100 40GB, 521GB, 2xNVME)
- 8.6 PiB Lustre workstorage, 290 TiB IME burstbuffer, 340 TiB SpectrumScale homestorage and 7 PB StorNext HSM tape archive

Storage systems at GWDG

- SCRATCH/WORK:
  - 8.6 PiB DDN Lustre (130 TiB SSDs)
  - 290 TiB DDN IME Burst Buffer
  - 3 PiB BeeGFS (200 TiB SSDs) in 4 instances
- HOME and software:
  - 340 TiB Spectrum Scale
  - 2.5+1.5 PiB StorNext several filesystems in 2 instances
  - 7 PiB NetApp filer
- Cold/PERM:
  - 20+ PB Ceph object storage
  - 35+ PB StorNext HSM tape archive in 2 instances

Problems

- Taking care of a large number of filesystems needs time
- Different filesystems have different kernel version requirements, intersection empty
- Filesystem usage between islands very uneven
- Heavy performance overprovisioning necessary for sufficient performance in each island
- Copies of files in different filesystems wastes storage space
- Educating users about the different filesystems is difficult

Challenges

- Neverending increase in space requirements
- Data analytics and machine learning require increasing amount of IOPS, because of non-sequential file access
- Users are spoilt by local SSDs in laptops and workstations
- R/W distribution changes from balanced (traditional HPC) to read intensive up to WORM.
- Agility of data processing workflows increases
- Sharing of data between data centers (in a secure way)
- Gap between active and inactive data becomes larger

Traditional HPC

- Starts with a model description
- Creates large simulation results
- Intermediate steps are written in large checkpoint files
- Meshes are getting finer or interactions involve more objects resulting in increasing requirements of compute ressources and larger reults
- Analyzes data often directly on the HPC system in post-processing jobs
- Final results have to be exported to home/partner institute or go directly to some type of archive
- Most intermediate results and checkpoints can be deleted

Data Analytics and AI

- Large datasets have to be ingested to the HPC system for analysis
- After initial data cleanup und preprocessing static dataset
- Read intensive
- Often random read
- Large number of small files
- Embarrasingly parallel jobs work on the same dataset $\rightarrow$ even more random reads and metadata hotspots in the namespace
- Results are often highly condensed statistics $\rightarrow$ small files
- Dataset and results have to be archived for reference
- Interaction with cloud services outside of HPC

Data Analytics with experimental data

- Data for analysis is created in lab experiments (1+ GiB per minute)
- After initial cleanup and manual preprocessing in the lab, data ingest to HPC
- Analysis with similar requirements like data analytics and AI case over night
- Result inspection
- Rework the manual preprocessing and ingest again
- Analysis and inspection
- Repeat until satisfaction in a highly competitive environment
- Processed data and results will be shared with cooperating institutes
- Raw data, processed data and results have to be archived
- For medical data take all of the above and add encryption requirements

Usability challenges

- How to automize the ingest, analysis, export workflow
- How to share the data in a secure way (server process running as root is not an option)
- Data sharing and migration between storage systems should have a unified interface $\rightarrow$ S3
- How can a sensible data management be implented, which has a minimal overhead for the user and is easy to understand
- How to achieve long term availability of data and results which have regulatory archival periods in a cost efficient way (data availability and metadata for finding the data again)? $\rightarrow$ Data Lakes with automated metadata extraction and enforcement of metadata supply for non-extractable information