The Software Defined Online Storage System at the GridKa WLCG Tier-1 Center

CHEP 2018, Sofia
Jan Erik Sundermann, Jolanta Bubeliene, Ludmilla Obholz, Andreas Petzold
GridKa Online Storage System

- New storage system in production since 2017
- Storage system design allows for scalability both in size and performance
- Spectrum Scale (v4.2) software defined storage
  - Few large file systems for better scalability and manageability
  - Allows to adapt to various use cases (4 experiments, tape buffer, …)
- Total capacity: 34PB
Technical Design

- NEC GxFS Storage Appliance with
- 22 redundant storage servers
- 90 disk enclosures
- 5400 HDDs for data (8+10TB)
- 58 SSDs (1.6 TB) in separate enclosures for file system metadata
- Two redundant IB fabrics (Mellanox 56G FDR)
- 64 protocol servers with 40GE
- 8x 100GE uplink to GridKa backbone
Storage Building Blocks (SBB)

- Each SBB with
  - 2 servers (NEC Express 5800 R120g-2M)
  - Redundantly connected to 2x5 disk enclosures (NetApp E5600 / DE6600)
  - 60 disks per enclosure (8TB / 10TB NL SAS)
- Setup in Dynamic Disk Pools (DDP) with
  - 50 disks per DDP
  - Data stored as raid6-stripes (8+2P)
  - Disk space corresponding to 3 disks as reserved capacity for fast rebuild
File Servers: xrootd

- File servers separated form storage in „remote“ protocol clusters
- Native xrootd setup for ALICE
- Each xrootd file server is able to serve all files
  - Directly profits from parallel file system
  - Easy to scale throughput by adding additional file servers
File Servers: dCache

- File servers separated from storage in „remote“ protocol clusters
- Separate dCache instances for ATLAS, CMS, LHCb, Belle II
- Each dCache pool lives in sub-directory of large file system
  - No automatic failover but easy to deploy pool on new hardware
  - Work in progress to improve pool startup times

```plaintext
Spectrum Scale Storage Cluster with file system

/exports/atlas/pool1
/exports/atlas/pool2
... 
/exports/atlas/poolx

Protocol Server Cluster

dCache 1
dCache 2
... 
dCache x
```
Deployment and Configuration

- Centralized host management for different infrastructures
  - All configuration in GitLab
  - Rolled out via continuous integration system
  - Puppet / Foreman host configuration & deployment

- Spectrum Scale integration
  - New nodes join existing Spectrum Scale clusters automatically
  - Reinstalled nodes recover Spectrum Scale config automatically
  - Nodes join existing CES clusters automatically
Monitoring and Performance Data

- Started to use ELK stack to collect log files from storage servers
- Helps to correlate errors events of cluster nodes
- Example: RDMA connection errors due to broken IB cable

- Collect performance and health data with Telegraf / InfluxDB / Grafana
Performance

- User requested maximum read/write performance so far: 50GB/s
- Usage statistics (2017 average rates per month)
  - 7.5 / 4.2 PB read from compute farm / remote
  - 1.1 / 3.0 PB written from compute farm / remote
Experience with Storage Expansion

- GridKa storage expansion for 2018/2019 has already been setup
  - Capacity: 23 → 34 PB
  - # protocol servers: 44 → 64, # NSD servers: 16 → 22
  - Expected combined r/w performance: 70GB/s → ~ 100GB/s

- Transparently included in existing IB fabrics and GPFS file systems

- During upgrade changed setup of IB fabrics from 1:1 → 1:2 blocking factor
Summary and Outlook

- New online storage system successful in production @GridKa WLCG Tier-1 center
- Transparently expanded system with 2018/19 pledges
- Started investigating technological options for time beyond 2019
- ~20% per year resource increase envisaged
  ➢ ~49PB disk storage till 2021
- Ready for scaling system for requirements towards HL-LHC