Karlsruhe Institute of Technology

Providing Reliable Storage – GridKa Tier 1 / T1_DE_KIT / FZK / FZK-LCG2 / LCG.GRIDKA.DE

GridKa Team

STEINBUCH CENTRE FOR COMPUTING - SCC

Outline

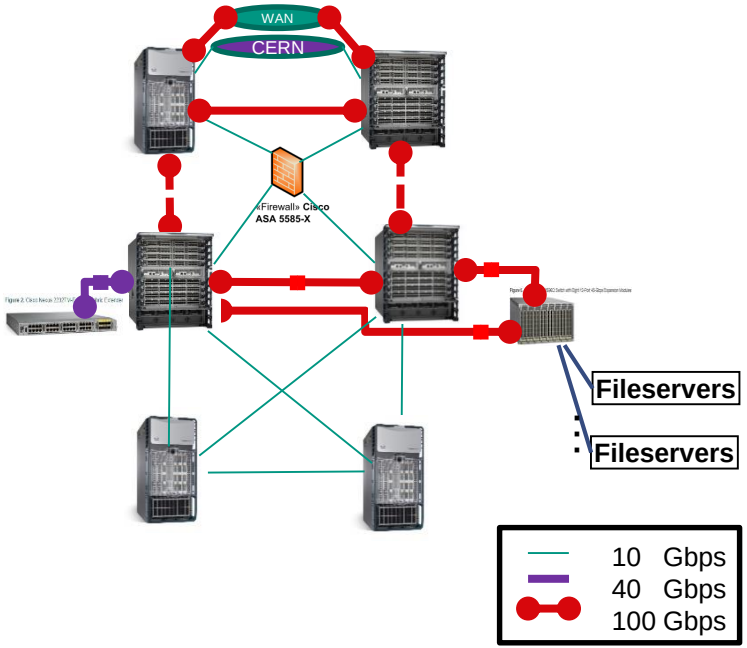
- Power & Cooling
- Network
- Storage Hardware
- GPFS
- Storage Elements

Power & Cooling

- All storage racks have dual power feeds (UPS and line power)
- Battery UPS
  - No generator etc. because no lives at stake
- Dual independent power feeds to KIT Campus North
- All storage and file server racks are water cooled
  - Cold water supplied by combined power/heating/cooling plant on campus
  - Old chillers used as backup
  - $10m^3$ cold water buffer

Network

- 2 border routers (Cisco Nexus 7010/7710) for redundant WAN connections
  - 2x 10G to CERN, 2x 100G to LHCONE/OPN/GPI
- 4 internal fabric routers (Nexus 7010/7710)
  - WNs racks each connected via fabric extenders to only 1 router (no redundancy)
  - Latest generation of file servers connected to Nexus 5696Q (TOR switches) with redundant uplinks to two Nexus 7710
  - Older file servers connected with 1x/2x 10G to one router (no redundancy)
- Our experience: router chassis rarely fail; transceivers and line cards fail; ISSU not 100% reliable



Storage Hardware

- Tier-1: ~25PB online storage (2017), ~32-35PB (2018/2019)
- 13 NEC SNA660 (i.e. NetApp E5600 series)
  - 300 8TB HDDs in 5 Enclosures, each 5 drawers of 12 disks
  - Redundant active-active controllers, redundant SAS connections to servers
  - 60 disks (2 per drawer) per dynamic disk pool with 3 drives reserved capacity
- 2 NetApp E2800 for GPFS Metadata
  - 29 1.6TB SSDs each
  - RAID1 with hot spare
- 1 DDN SFA12K
  - To be retired by end of 2018

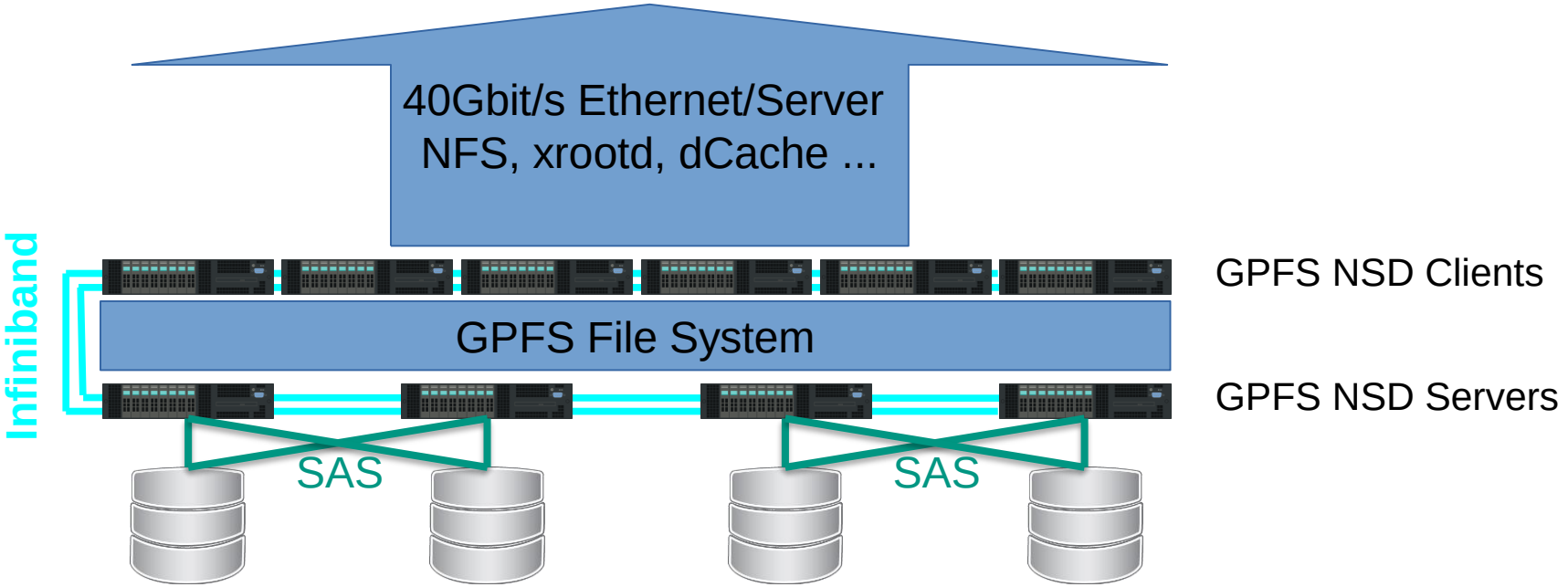
- Our strategy: prevent disk/drawer/enclosure/controller failures from being visible on the file system level → storage hardware setup needs very good failure handling

IBM Spectrum Scale (GPFS)

- Software defined storage (long before people were using that term)
- Support for
  - Very large file systems, with large number of files
  - Storage class tiering
  - Simple online scaling of capacity
  - Transparent storage hardware migration
  - Redundant disk connections
  - Separate data/metadata handling
  - No metadata bottlenecks by design
  - POSIX
  - Nowadays also integrated NFS/CIFS/S3 services

GPFS at GridKa (latest storage generation)

- 4 copies of the metadata
  - 2 copies by GPFS
  - 2 “copies” by RAID-1 of metadata storage systems
- 1 (or 2) filesystem per VO
  - ATLAS 8.6PB, ALICE 5.9+0.66PB, CMS 5.3PB (LHCb uses different storage systems)
  - Lesson from the past: segregate VO workloads
- NSD server cluster
  - Hosts all file systems
  - 16 servers: each NSD is accessible via two servers
- 1 “protocol server” cluster per VO (8-10 servers)
  - GPFS remote mount of VO file system
  - Separation of GPFS (NSD) servers and dCache/xrootd servers


40Gbit/s Ethernet/Server
NFS, xrootd, dCache ...
Infiniband
GPFS File System
GPFS NSD Clients
GPFS NSD Servers
SAS
SAS

dCache at GridKa

- Separate dCache instances for each VO (currently 4 production, 1 test)
  - Separation of VO workloads
  - VO specific tunings
  - Easier maintenance
- Resiliency
  - 2 gridftp/webdav/xrootd doors per instance
  - Master-slave postgres setup, but manual switchover, not HA
  - Currently not exploiting dCache high-availability features yet – 1st candidate SRM door
- dCache pools can’t share their inventory, so no direct profit from GPFS
  - If a pool server fails, pools can be started on other servers, because all data is on GPFS

Xrootd for ALICE

- Two independent xrootd redirectors
- Xrootd data servers share all data on one filesystem
  - Every server can serve any file on disk
  - Profits directly from GPFS