Karlsruher Institut für Technologie

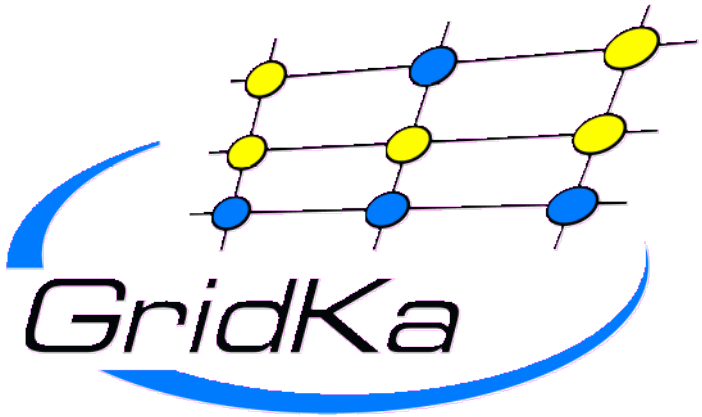
# KIT Site Report


## Andreas Petzold

STEINBUCH CENTRE FOR COMPUTING - SCC

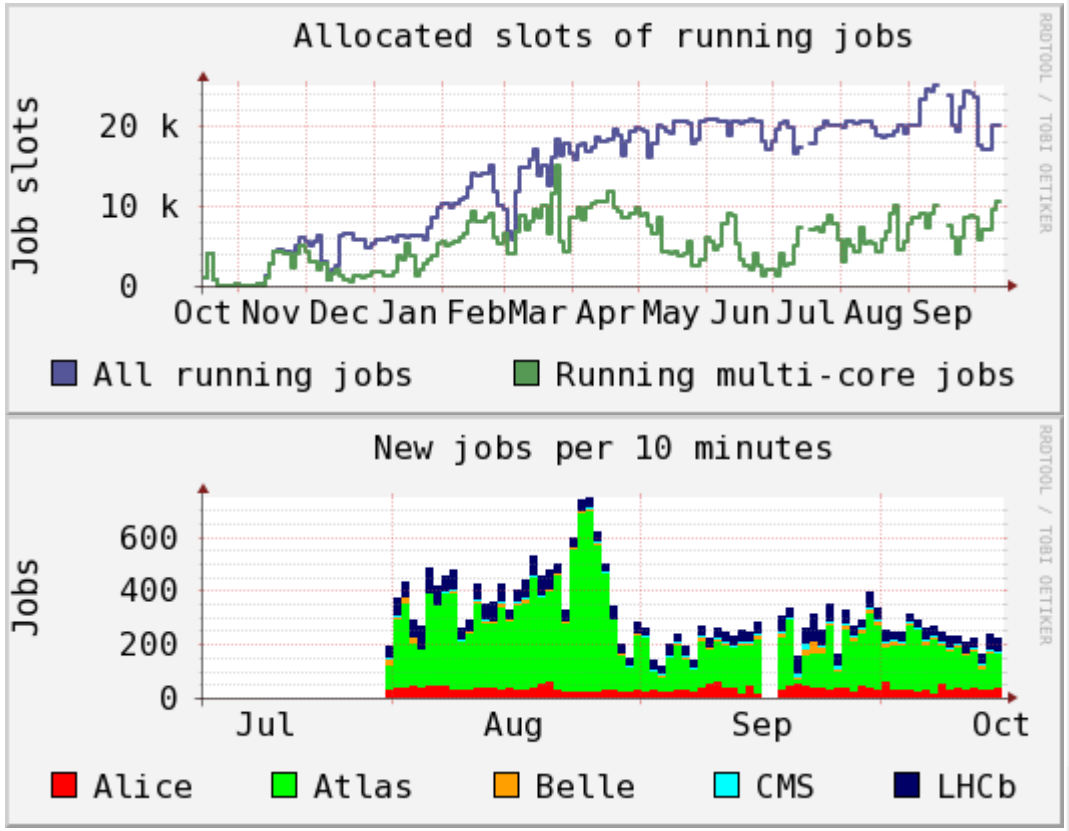

# Outline

- Batch System & WNs
- Tape
- Network
- Disk Storage



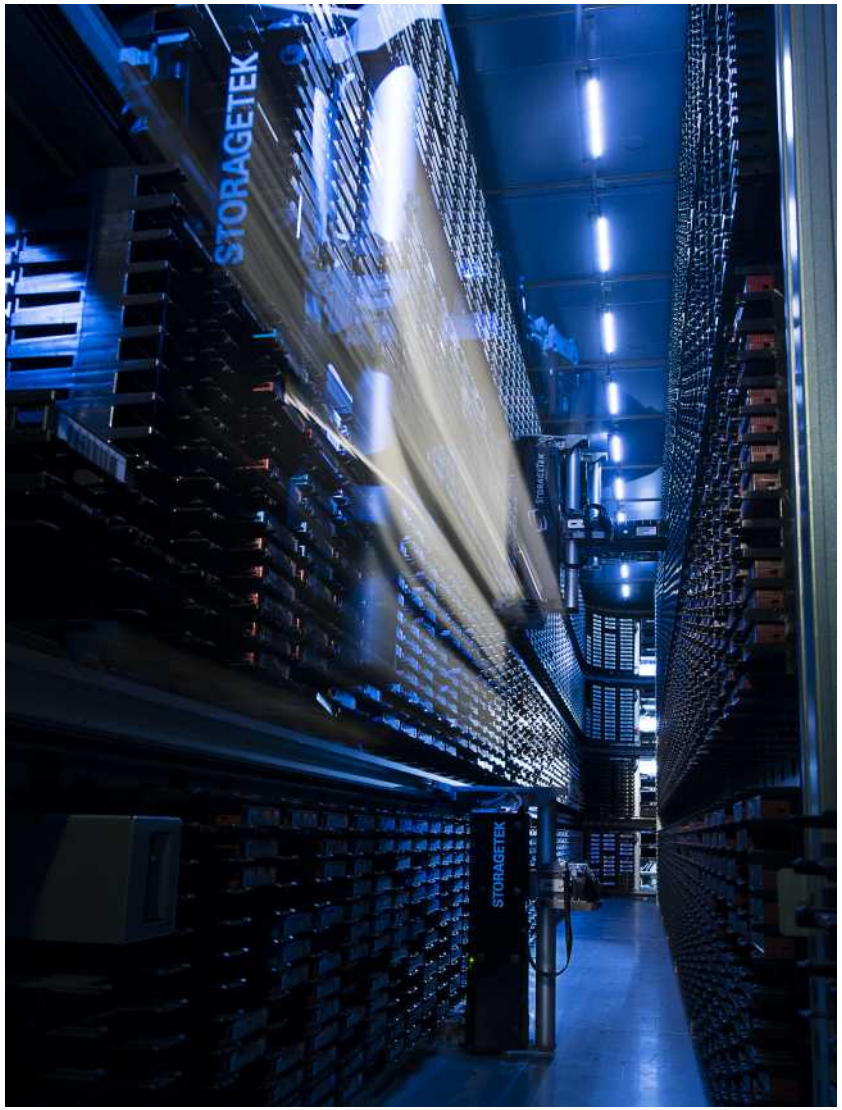

# Tier-1 Batch System & CEs & WNs

- HTCondor only since March 2017
  - smooth operations, but we had a lot to learn
  - continuous optimization of MC/SC scheduling and defragmentation
  - problem w/ empty ATLAS pilots → OK with aCT
- ARC CEs
  - many stability and scaling issues until summer
  - patches required for ALICE and LHCb
- Worker Nodes
  - 850 Nodes, 320kHS06, ~24000 SC job slots
  - 1.5/1.6 job slots per core
  - now only Intel except for few AMD test systems



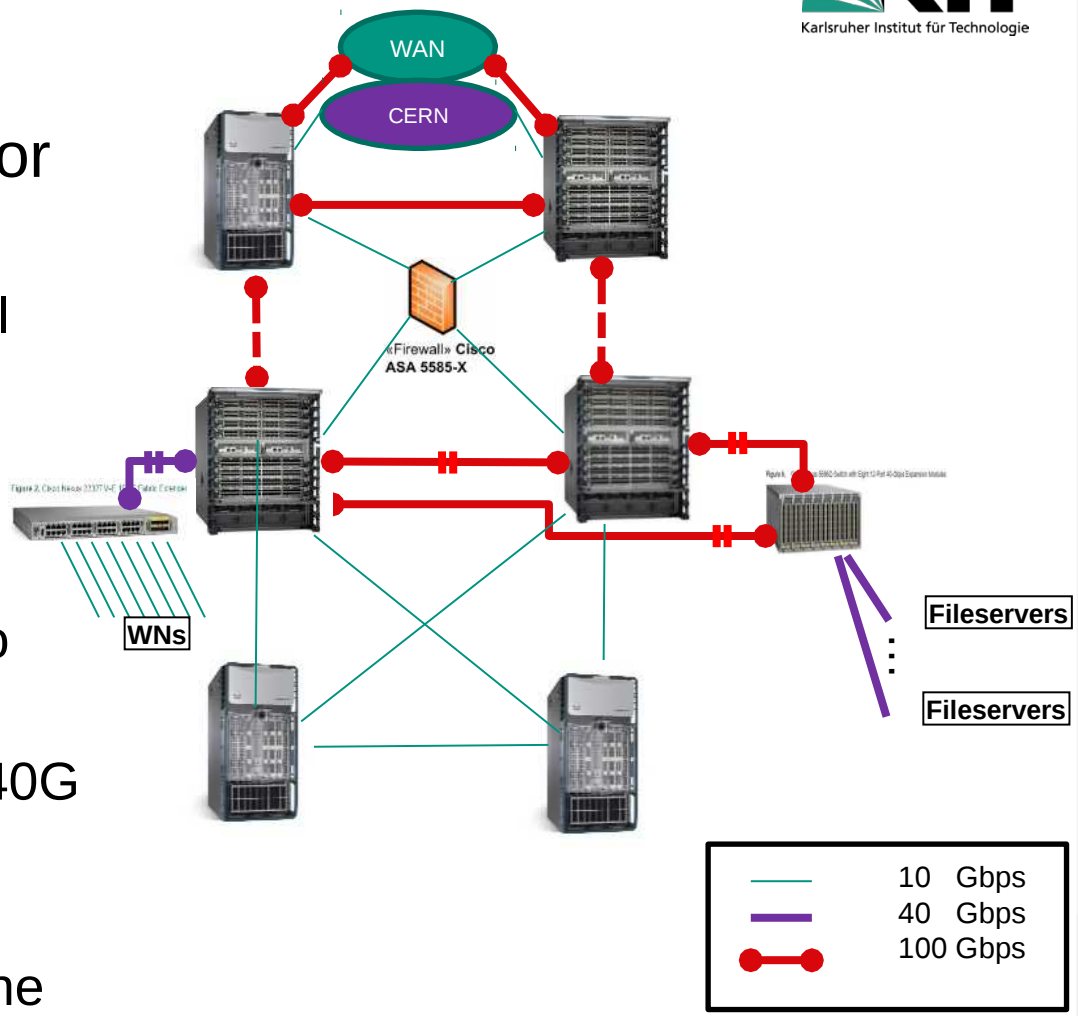
# Tape

- 2 Oracle SL8500, 2 IBM TS3500
  - 20 T10K-D, 17 T10K-C, 5 TS1140, 5 LTO-5
  - several problems with T10K-C drives
    - “insufficient write efficiency”, ”stuck roller”
    - O(1000) cartridges affected
    - only monitored by latest firmware builds
- Tier-1
  - still on TSM
  - preparing HPSS migration
- LSDF
  - switching from TSM to GPFS-HPSS-Interface (GHI) for disaster recovery



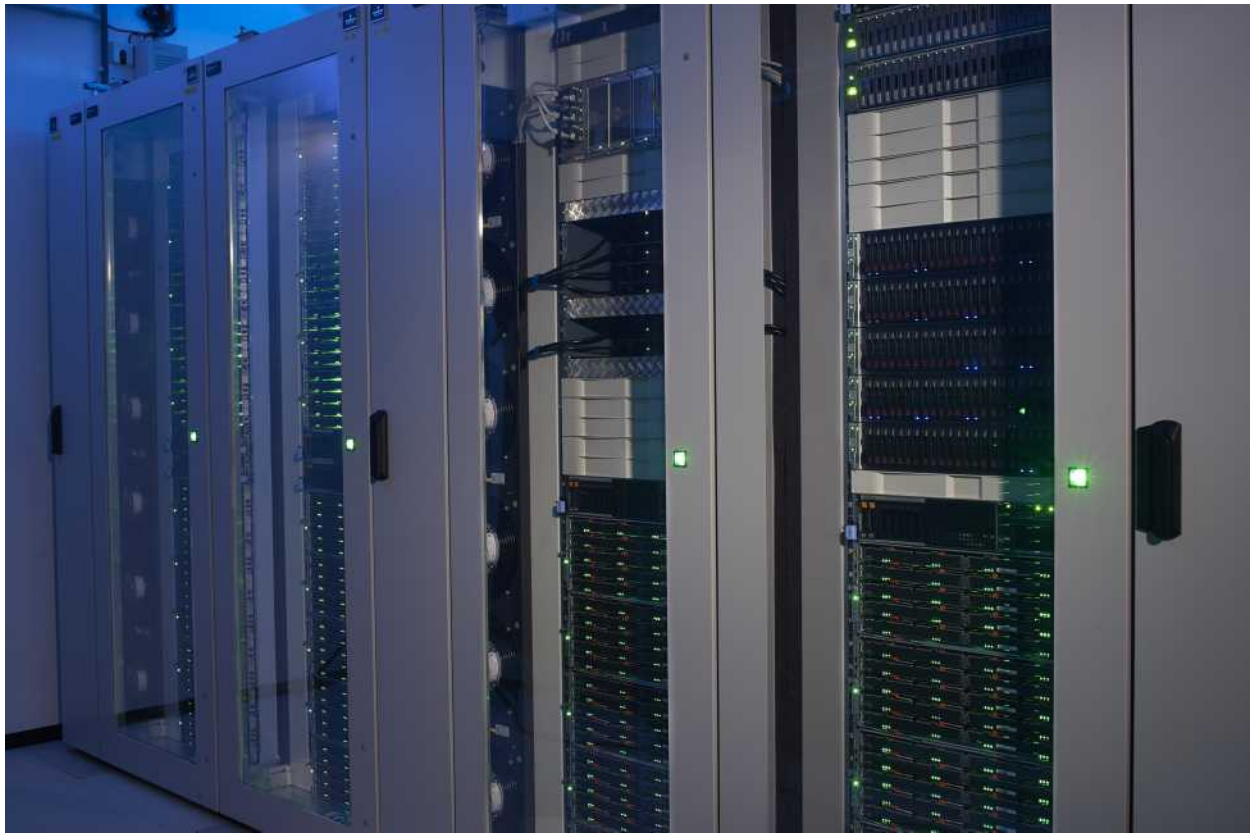
# Network

- 2 border routers (Cisco Nexus 7010/7710) for redundant WAN connections
  - 2x 10G to CERN, 2x 100G to LHCONE/OPN/GPI
  - 100G(+20G backup) to CERN planned for 2018
- 4 internal fabric routers (Nexus 7010/7710)
  - WNs have 10G
  - WN racks each connected via fabric extenders to only 1 router (no redundancy) with 40G
  - Latest generation of file servers connected with 40G to Nexus 5696Q (TOR switches) with redundant 4x100G uplinks to two Nexus 7710
  - Older file servers connected with 1x/2x 10G to one router (no redundancy)



# Online Storage System

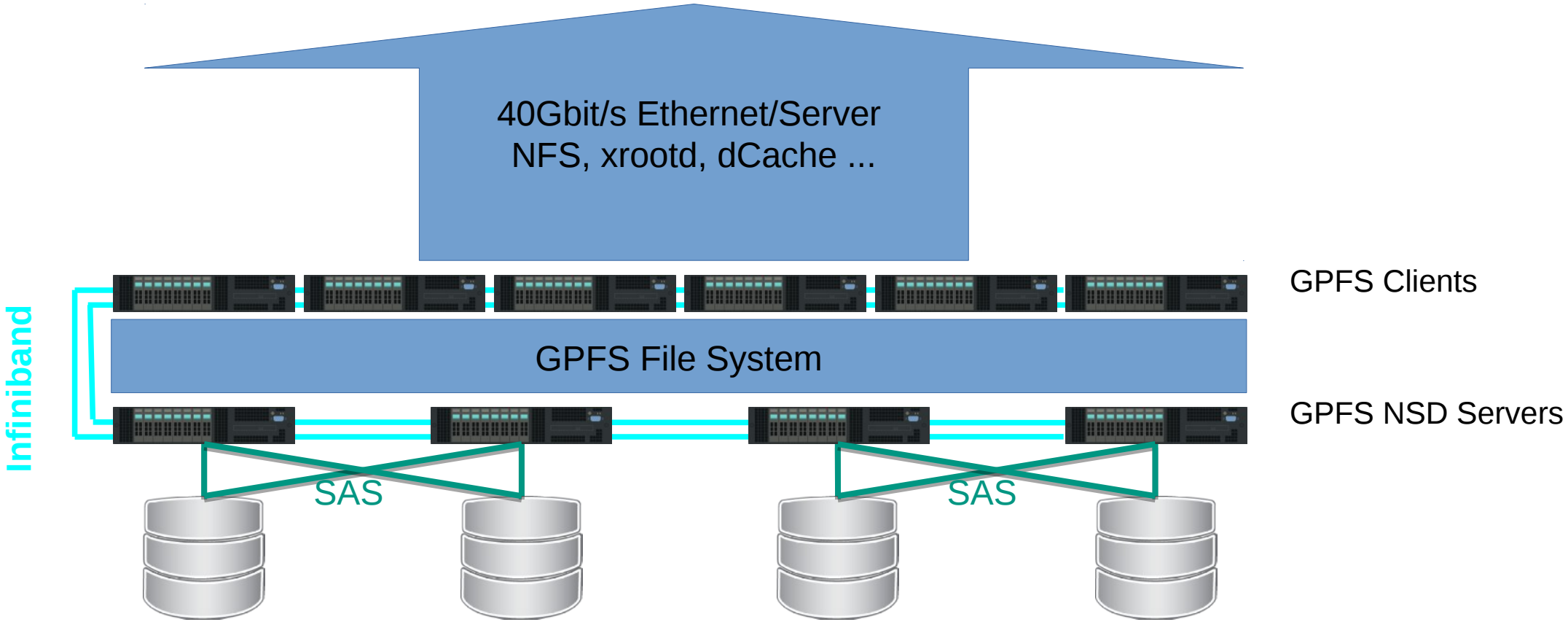
- new storage systems installed in late 2016
  - 20PB for GridKa, 6PB for LSDF
- in production since April 2017
- 2.7+2PB extensions installed late summer
- 11+2.2PB extensions to be ordered
- NEC as system integrator
  - NEC (OEM NetApp) storage systems
  - IBM Spectrum Scale (aka GPFS)
  - NEC GPFS NSD servers
  - NEC (OEM SuperMicro) protocol servers
  - Mellanox IB fabric



# Storage Systems Hardware

- 23PB system for Tier-1
- 13 NEC SNA660 (NetApp E5600)
  300 8TB HDDs each
  2 redundant active-active controllers,
  60 disk DDPs instead of RAID-6
- 2 NetApp E2800 for GPFS Metadata
  29 1.6TB SSDs each
  2 redundant active-active controllers, RAID-1
- 16 NSD servers (20 cores, 128GB RAM)
  SAS to disk controllers, FDR IB to protocol nodes
- 12 Mellanox FDR IB Switches in 2 redundant non-blocking fabrics
- 44 protocol servers (20 cores, 128GB RAM) for NFS/dCache/xrootd
  FDR IB to NSD servers, 40G Ethernet to WNs/WAN
- Benchmark Throughput: 70GB/s write+read

# Online Storage System Layout



# GPFS Setup

- 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  NSD servers and dCache/xrootd servers

# dCache/xrootd on GPFS

- 1 dCache pool per server / 8-10 servers per VO
  - much better tuning of mover limits (currently almost none)
  - challenges: ATLAS ~1PB/60M files per pool → ~6h untils pools becomes r/w
- 1 file system per xrootd SE
  - allows us to deprecate seperate namespace directories (oss.space+localroot)
- Data migration
  - Phase 1: automatic via dCache of ATLAS/CMS/BelleII; ~6.5PB moved in ~1 month
  - Phase 2: rsync of ALICE xrootd data; ~4.5PB; very slow, due to file size
  - Phase 3: automatic via dCache of LHCb to DDN SFA12K based storage to start next week
- 13.5PB of old DDN storage is/will be switched of by end of 2017

# Thank You!