Karlsruhe Institute of Technology

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

Site Report – Karlsruhe Institute of Technology (KIT)

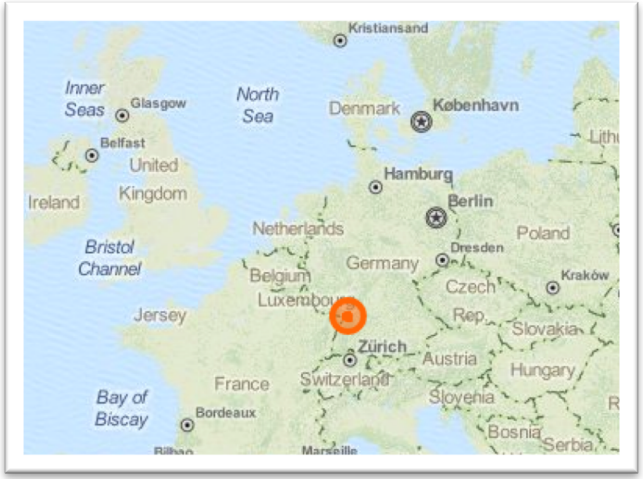
UK 2018 Spectrum Scale User Group Event, London

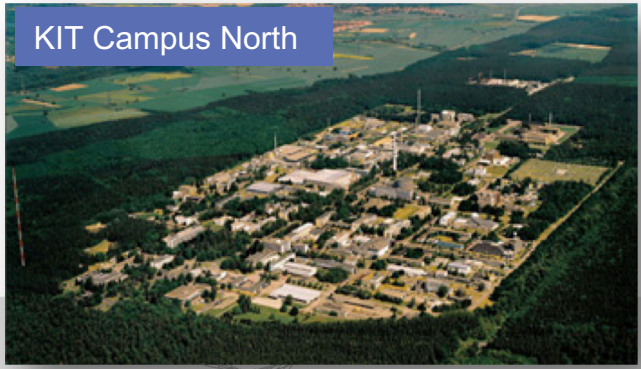
Jan Erik Sundermann

STEINBUCH CENTRE FOR COMPUTING (SCC)

Introduction

- Karlsruhe Institute of Technology (KIT)
  - State university with research and teaching and
  - Research center of the Helmholtz Association
  - 25000 students, 9500 employees,
  - 844M€ annual budged in 2013
- Steinbuch Center for Computing (SCC)
  - Founded on January 1st, 2008
  - Two locations at KIT Campus South and North
  - ~200 people in total
    - 60% scientists, 40% technicians, administraion
  - 7 departments and 4 research groups





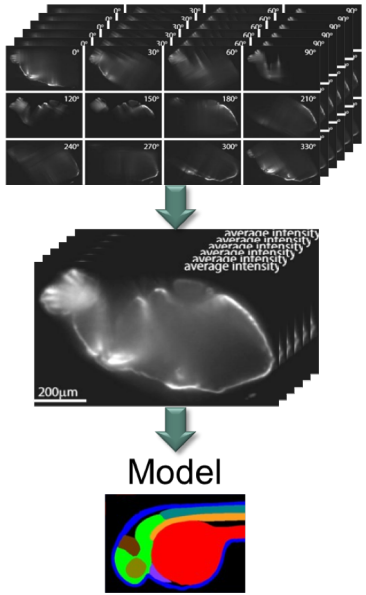
Enabling Data Intensive Science at SCC

- SCC at KIT operates large storage infrastructures for data intensive science:
  - Grid Computing Centre Karlsruhe (GridKa)
    German Tier-1 center for the world wide LHC computing grid
  - Large Scale Data Facility (LSDF)
    Multi-disciplinary storage infrastructure for data intensive science disciplines
  - Smart Data Innovation Lab (SDIL)
    National R&D platform for Big Data with Industry and SMEs

Large Scale Data Facility

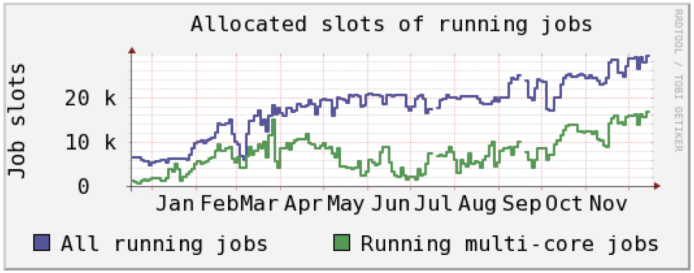
- Storage facility for diverse set of fields of science
  - Climatology, photon science, structural biology, hydrodynamics, engineering, …
  - State service bwFileStorage for HPC users
- Migration to new storage systems almost complete
  - 10PB online-storage, 4PB offline storage
  - ~3 PB migrated from old storage using AFM
- 100Gb/s connection to partner installation in Heidelberg
- Variety of storage protocols: NFS, CIFS, SFTP
- Variety of connected clients
  - Experimental equipment, user desktops, clusters
  - Close connection to KIT and state wide HPC clusters via data mover nodes → direct GPFS access from HPC nodes planned
  - Plan to use GHI / HPSS for backup

Example ITG Image Processing



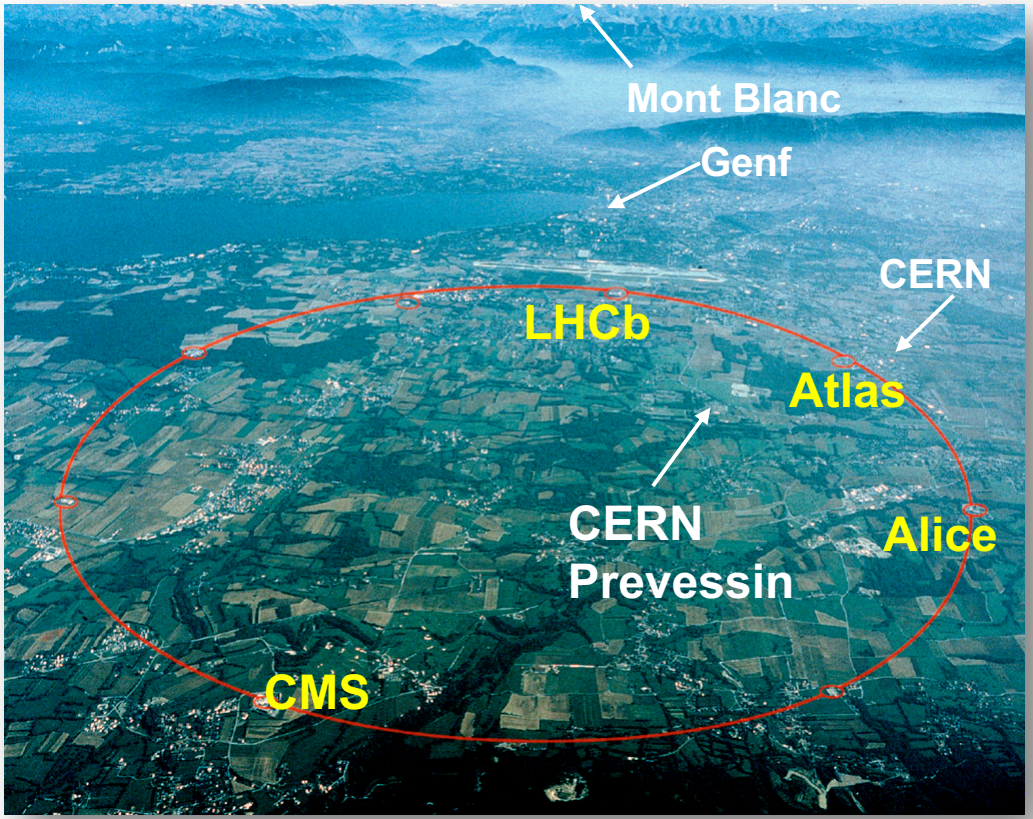
Grid Computing Centre Karlsruhe

- Data and computing center for particle and astroparticle physics experiments
  - German Tier-1 center in the Worldwide LHC Computing Grid (13 centers in total)
- High-throughput compute farm
  - 29k Cores, 32k job slots, 1100 worker nodes
  - Last 12 months: 24M jobs, 176M CPU hours
- Storage
  - 25PB on disk, 40PB on tape in 2 libraries
- Network
  - WAN: 20 + 100Gb/s to CERN and DFN
  - LAN: 80-200Gb/s backbone, storage servers 40Gb/s, WNs 10Gb/s

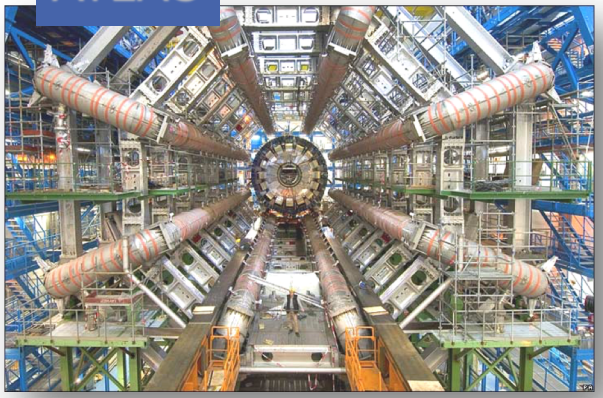


The Large Hadron Collider (LHC)

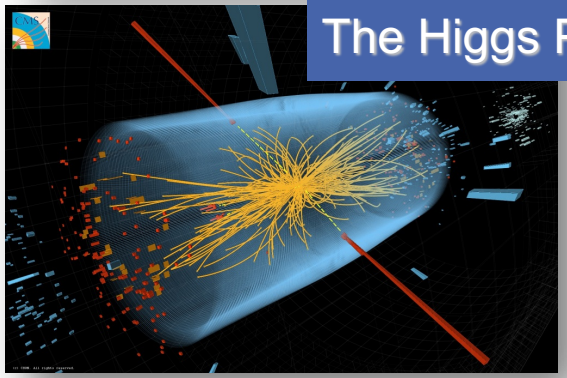
Karlsruhe Institute of Technology



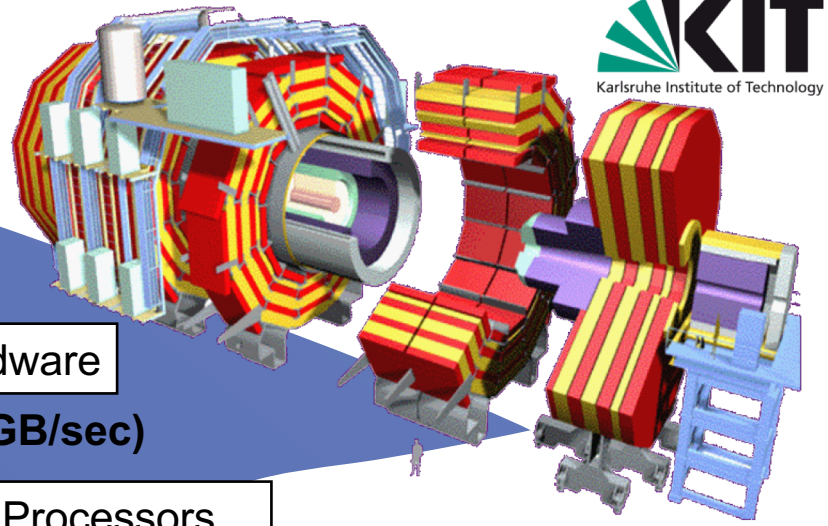
ATLAS



The Higgs Particle



HEP Data Acquisition



40 MHz x 25 MB = 1 PB/sec = 1000 TB/sec equivalent)

Data-reduction 1/10 Mio.

Level 1 - special hardware

75 KHz (75 GB/sec)

Level 2 - Embedded Processors

5 KHz (5 GB/sec)

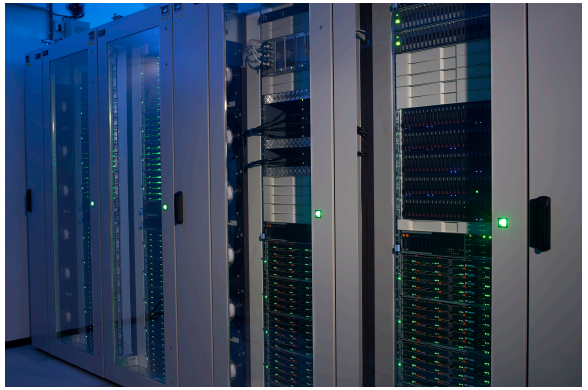
Level 3 – PC Farm(Linux)

100 Hz (100 MB/sec)

~ 2 PB per year per experiment RAW data (+ Simulations)

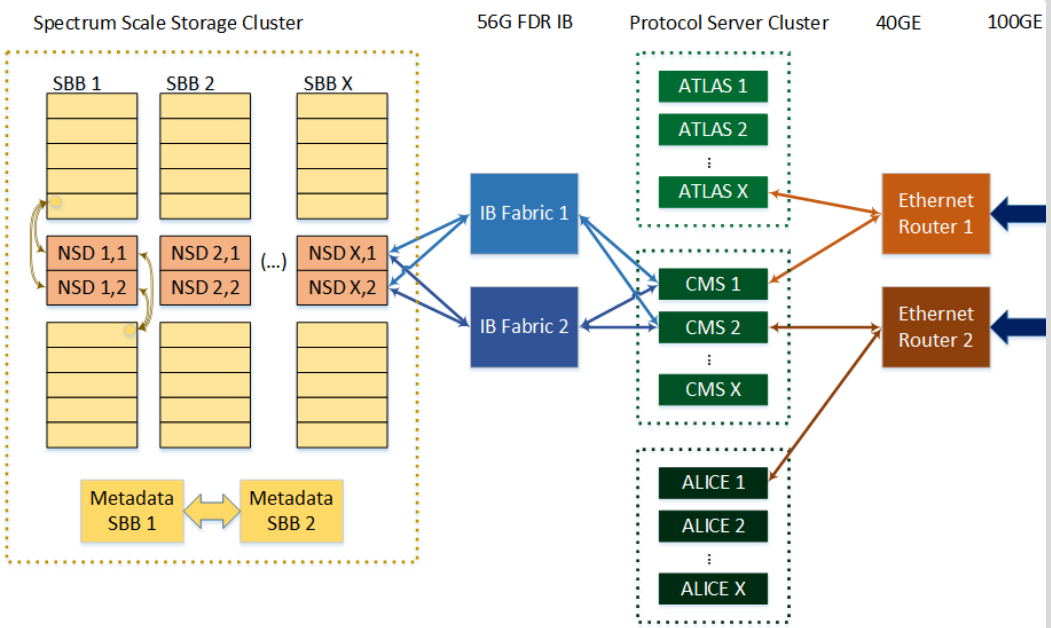
GridKa Online Storage System – NEC GxFS Storage Appliance

- 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: 23PB
- Other large infrastructures for scientific data @KIT (e.g. LSDF) follow same approach



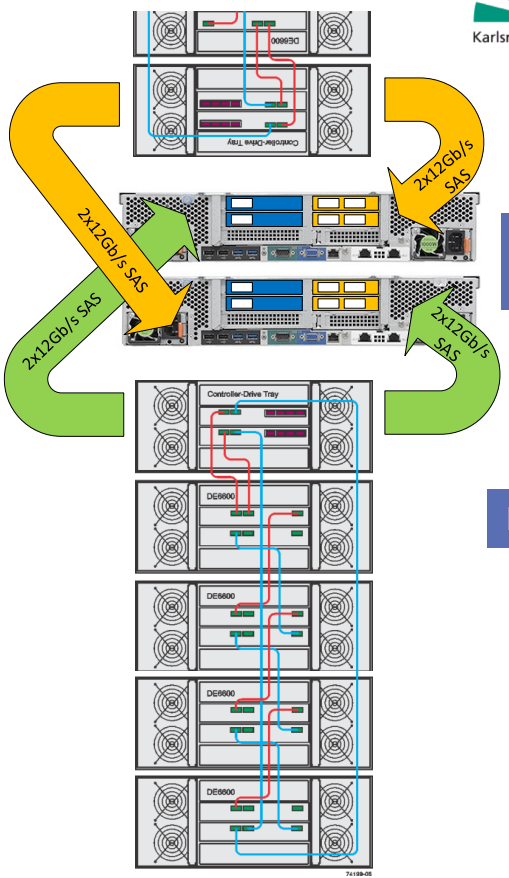
Technical Design

- NEC GxFS Storage Appliance with
- 16 redundant storage servers
- 70 disk enclosures
- 3900 HDDs for data
- 58 SSDs (1.6 TB) in separate enclosures for file system metadata
- Two redundant IB fabrics (Mellanox 56G FDR)
- 44 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

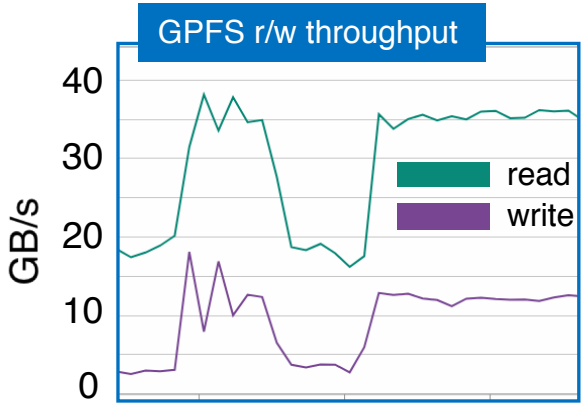


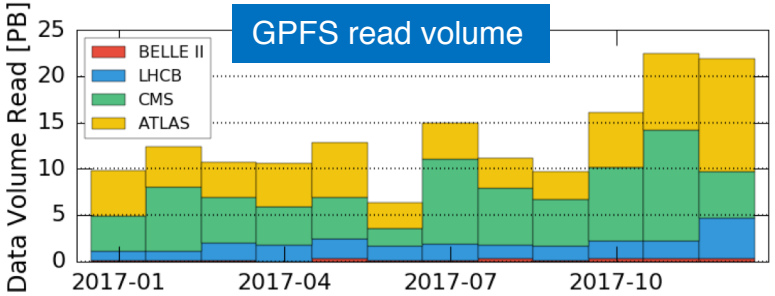
Protocol Servers

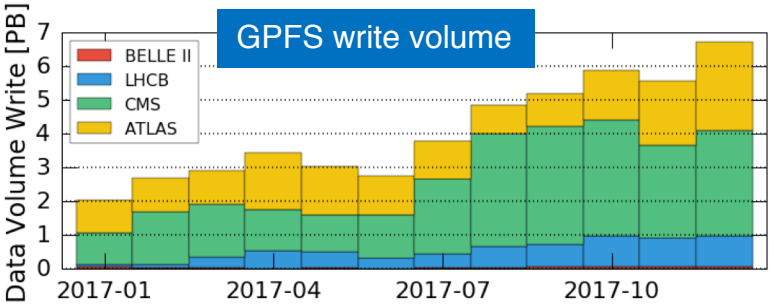
- File servers separated form storage in remote protocol clusters
- Mainly community specific protocols in GridKa
  - SRM, webdav, dcap, xrootd, gridftp, NFSv4.1
- LSDF and GridKa both with HA CES setup (NFS and CIFS)
- Custom CTDB setup for ssh/sftp user access in LSDF

Performance (GridKa)

- Benchmark r/w performance: 70GB/s
- User requested maximum read/write performance so far: 50GB/s
- Usage statistics (average rates per month)
  - 7.5 / 4.2 PB read from compute farm / remote
  - 1.1 / 3.0 PB written from compute farm / remote







Outlook

- GridKa storage expansion for 2018/2019 currently being setup
  - Capacity: 23 → 34 PB
  - # protocol servers: 44 → 64, # NSD servers: 16 → 22
  - Expected combined r/w performance: ~ 100GB/s
- Transparently included in existing IB fabrics and GPFS file systems
- ~20% per year resource increase envisaged
  - ~49PB disk storage till 2021
  - Further growth supported by capacity and performance scaling with IBM Spectrum Scale