Karlsruhe Institute of Technology

Komplexe HSM Systeme in Hochenergie-Experimenten am Karlsruhe Institut für Technologie

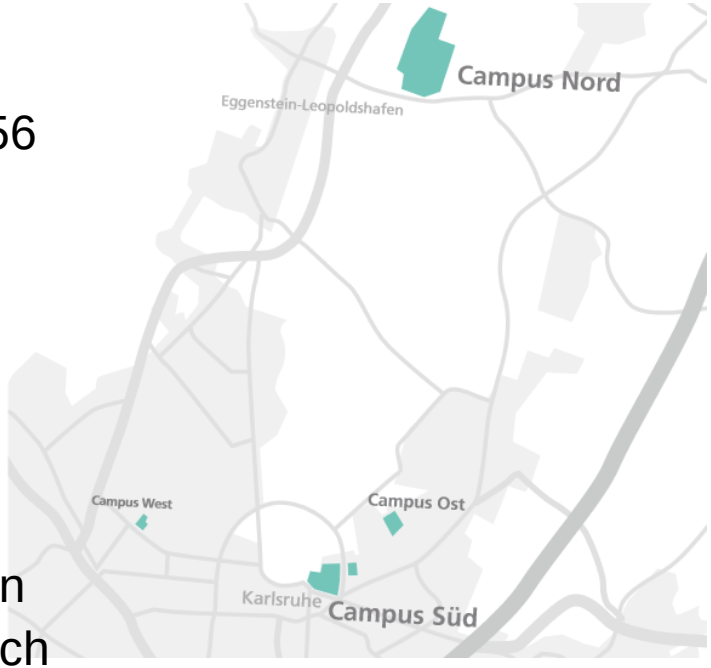
Dorin-Daniel Lobontu
Doris Ressmann
Preslav Konstantinov
Karin Schäfer

Agenda

- Karlsruhe Institut of Technology
- Steinbuch Centre for Computing (SCC)
- Tape related tasks
- GridKa – data and facts
- HPSS monitoring
- Migrating GridKa from IBM Spectrum Protect to HPSS

Karlsruhe Institute of Technology

- KIT founded 1. October 2009
  - Forschungszentrum Karlsruhe founded in 1956
  - University Karlsruhe founded in 1825
- Nearly 10.000 employees
- 22.373 students
- 368 apprentices
- Locations:
  - Campus North (Eggenstein-Leopoldshafen)
  - Campus South (University ground)
  - Campus East (Mobility campus)
  - Campus West
  - Dresden
  - Garmisch
  - Helmholz-Institut Ulm



Steinbuch Centre for Computing

- The Information Technology Center of KIT
- Center for data-intensive computing
- The analysis of large-scale data with high national and international visibility
- An innovative and agile IT service provider at KIT
- 300 employees from 25 nations in 16 departments and research groups

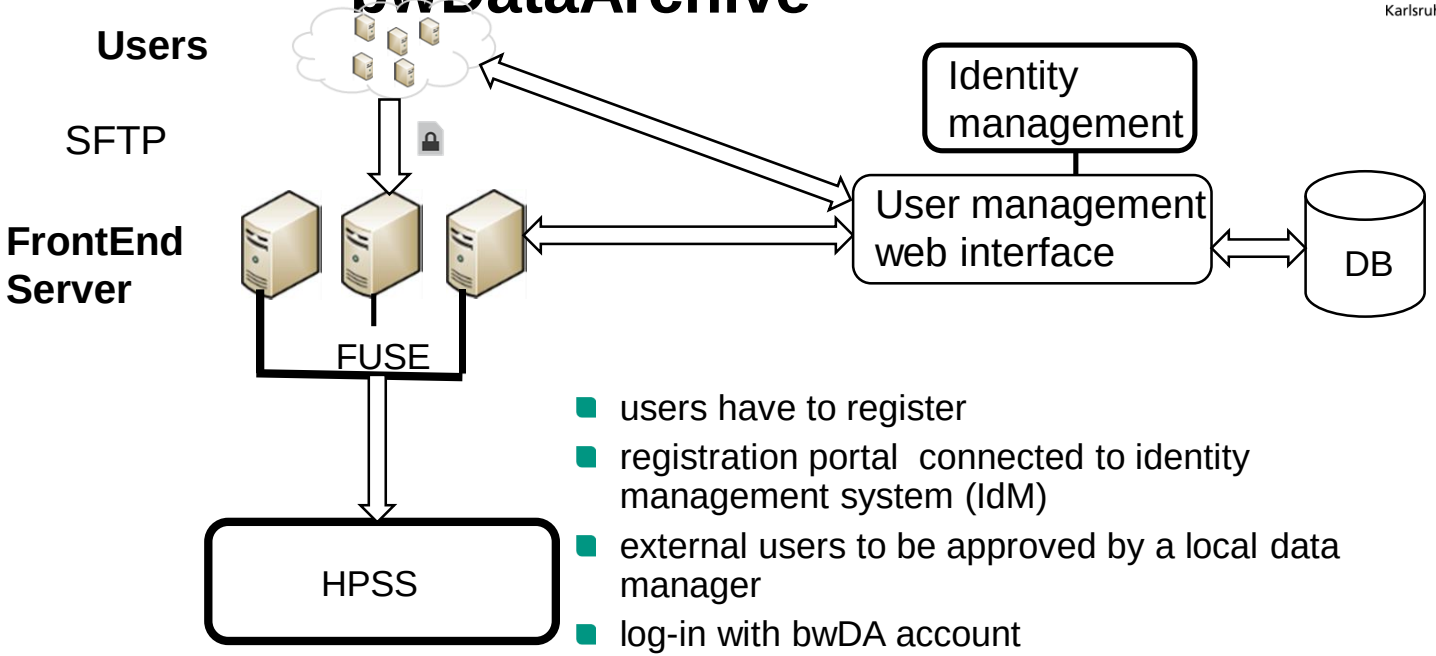
Tape related tasks

- KIT Backup via IBM Spectrum Protect (formally known as TSM)
- bwDataArchive uses High Performance Storage System (HPSS)
- Large Scale Data facility 22 PB in IBM Spectrum Scale (GPFS)
  - Classical Backup via IBM Spectrum Protect
    - Triggers full backup because of small metadata changes
    - Results in Backup duration of several days
  - Backup via GPFS HPSS Interface (GHI) in preparation
- GridKa the German Tier 1 Center of the Large Hadron Collider (LHC)
  - Tape backend used to be IBM Spectrum Protect
  - Migration to HPSS should be finished end of this year

bwDataArchive

- **Motivation:**
  - long- and safety-term preservation of data from scientific experiments, measurements, analysis and simulations
  - central and flexible system
    - unify all isolated islands of data
    - well scalable in terms of data transfer and amount of data
- **Goals:**
  - to offer an easy-to-use system to a non IT-confident scientific community
  - to identify and implement the safety-related aspects of a long-term storage
    - background data integrity verification
    - end-to-end data protection
      - avoid data corruption at the transfer time

bwDataArchive



- users have to register
- registration portal connected to identity management system (IdM)
- external users to be approved by a local data manager
- log-in with bwDA account
- transfer of data using SFTP
- transparent transfer to HPSS via FUSE mount point

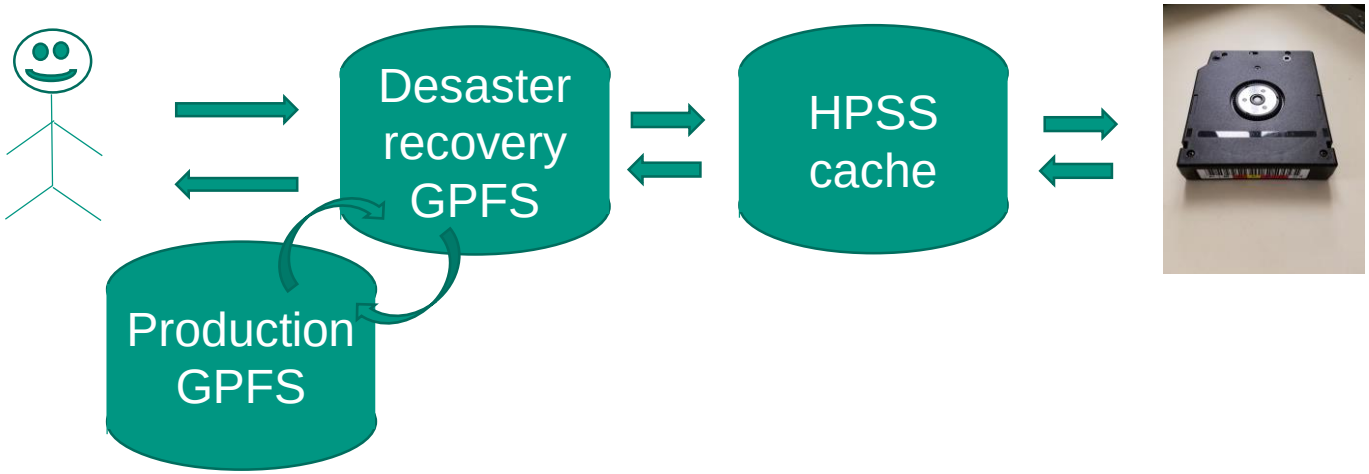
GPFS HPSS Interface (GHI)

- 22 PB GPFS space to Backup
- In ISP small metadata change triggers full backup
  - Results in Backup duration of several days
- GHI should not have such behavior
- GHI first setup
  - HPSS test system, GPFS production system
  - Dependencies between update cycles (HPSS, GHI and GPFS)
- GHI second setup
  - Productive GPFS -> disaster Recovery GPFS (fewer disk space with HSM functionality to HPSS)

Disaster Recovery GPFS System

- Normal operation
  - User access to production system
  - Transparent to user: migration to disaster recovery system
  - Transparent to user: HSM like migration to HPSS
  - Disaster recovery system holds only highly accessible data

Update Production GPFS



- Limited user access to disaster recovery system
- Access times might be slow if data has to be restored
- After downtime the production system will be synced again

Update HPSS

- Normal operation for users
- Connection to disaster recovery system will be stopped
- HPSS Update independent to the user
- After the update sync between production and disaster recovery system

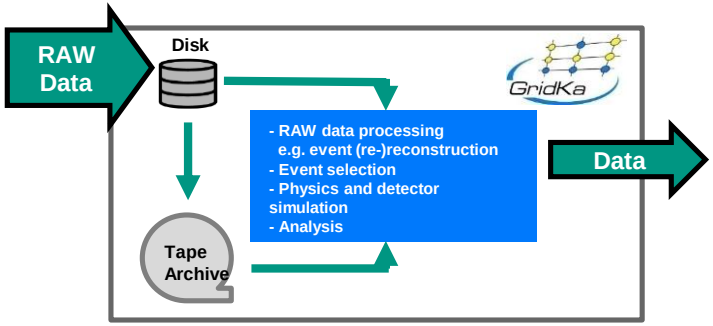
Gridka

- **Data and analysis center for particle and astroparticle physics**

PIERRE AUGER OBSERVATORY · BABAR · Belle II · ALICE · CMS · LHCb · WLCG Worldwide LHC Computing Grid ← ~10 000 physicists

- A **cornerstone** of the Worldwide LHC Computing Grid (WLCG)
- **Integral part** of the LHC data processing chain

Conclusion slide of R.D. Heuer, July 4, 2012



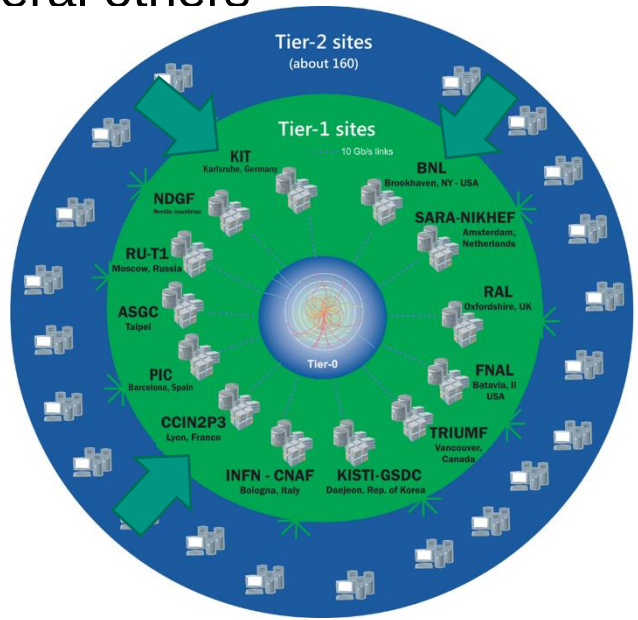
GridKa Building Blocks

- **20 year** nation Tier-1 Centre for LHC
- GridKa-Ressourcen in 2023
  - **~ 61.000 CPU cores**, 56 GPUs
  - **99 PB Online Storage** (6500 HDDs)
  - **135 PB Offline Storage** (Tapes)
  - **400 Gbit/s Network connection** (2x100 to CERN + 2x100 to DFN)
- GridKa in **global Scale**
  - **~15%** all Tier-1 CPU, Disk & Tape Ressource **worldwide** in WLCG

- *Computing at SCC is **much** more…*

GridKa – LHC Tier 1

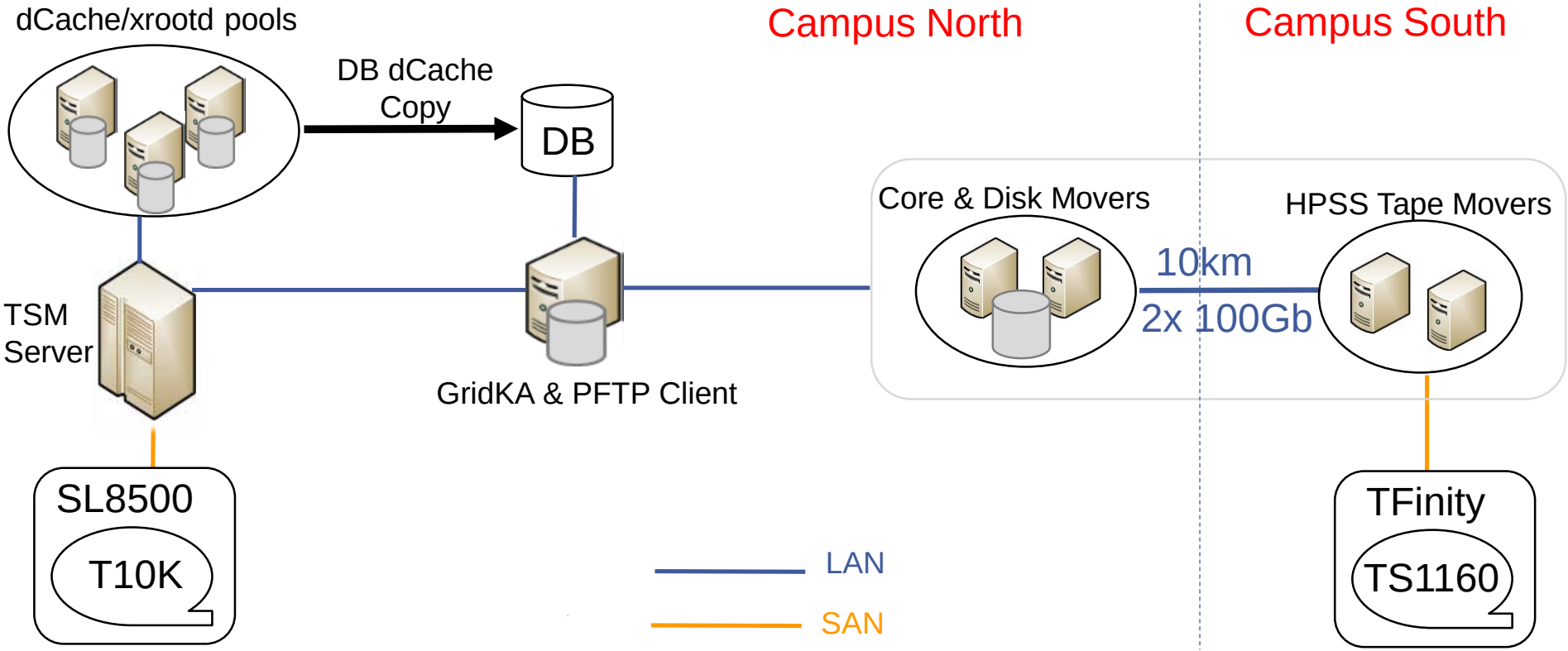
- support for particle physics computing
- stakeholders: 4 LHC experiments and several others
- Currently stored data: 70 PB, 392 Mil files



Worldwide LHC Computing Grid (WLCG)

- **~750k CPU cores**
- **600PB storage**
- **> 2M jobs / day**
- **10-100Gb links**

HPSS GridKa Layout



Offline Migration ISP to High Performance Storage System (HPSS)

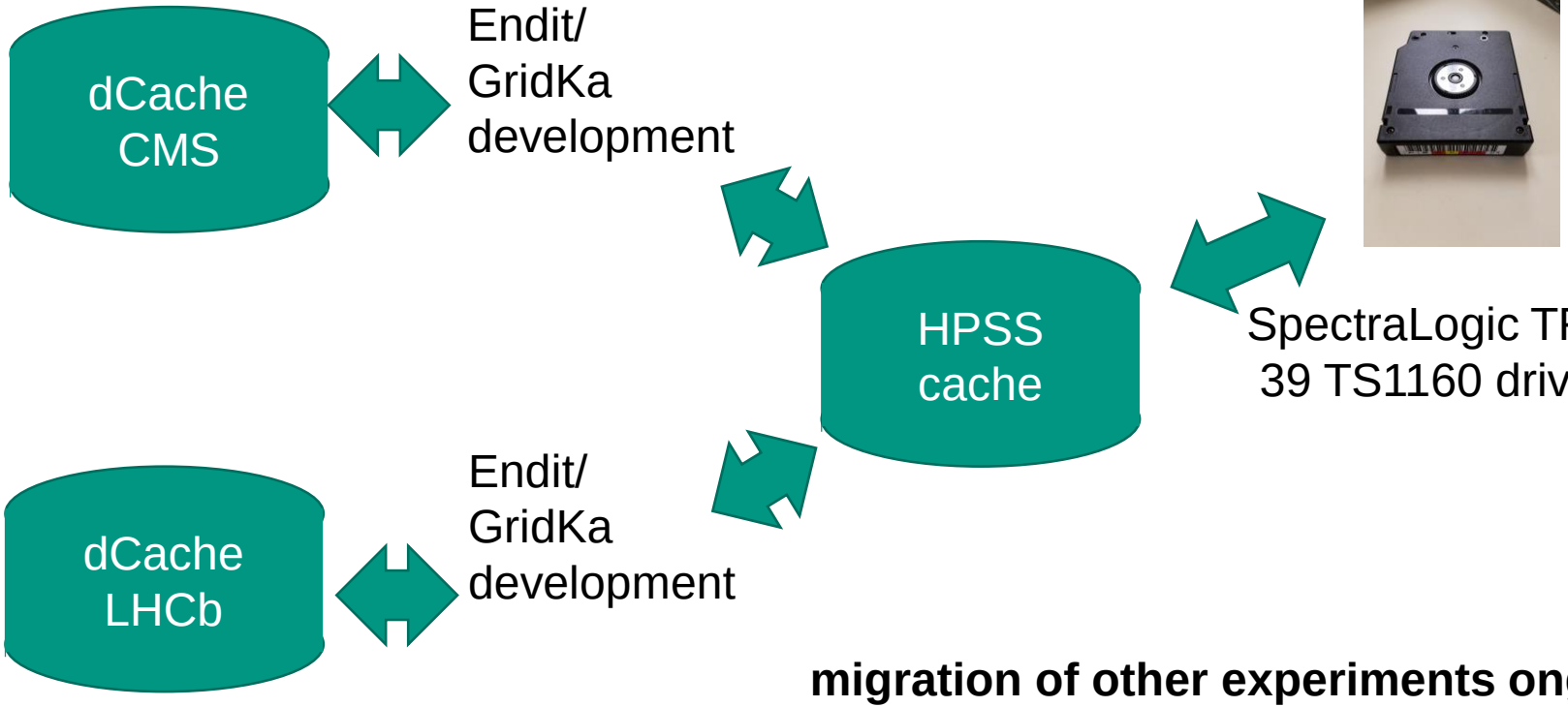
Long term team effort started in 2020

Dataset after Dataset

1 Oracle SL8500 Library
35 T10k-D drives

SpectraLogic TFinity
39 TS1160 drives

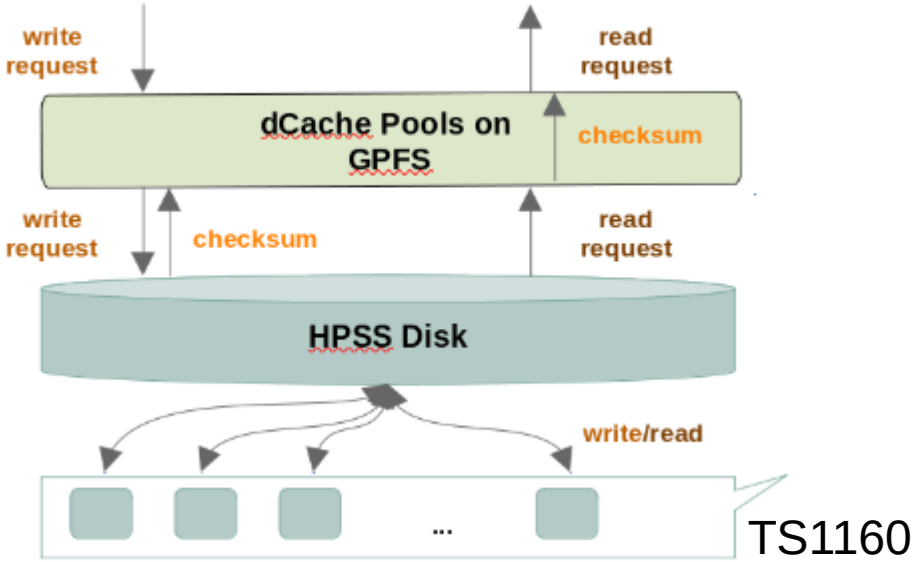
New Tape Connection
High Performance Storage System (HPSS)



migration of other experiments ongoing

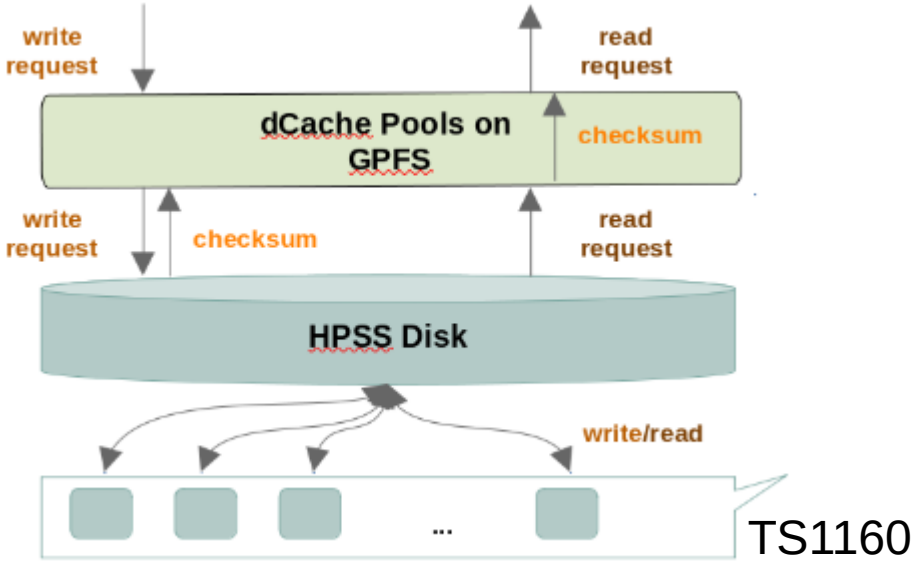
Writing files to tape

- Files are transferred from dCache pool providing checksum
- Written to HPSS disk buffer
- Checksum verification
- HPSS: tape writes are initiated in file aggregates by directory order
- Up to 100 files <= 10 GiB in one directory collected in aggregates



Reading files from tape

- Files read requests collected for file aggregates
- Using full aggregate recall mechanism (FAR)
- Files are read from HPSS disk buffer into the dCache pool
- Checksum verification is done by dCache



Migrate GridKa data from TSM to HPSS setup

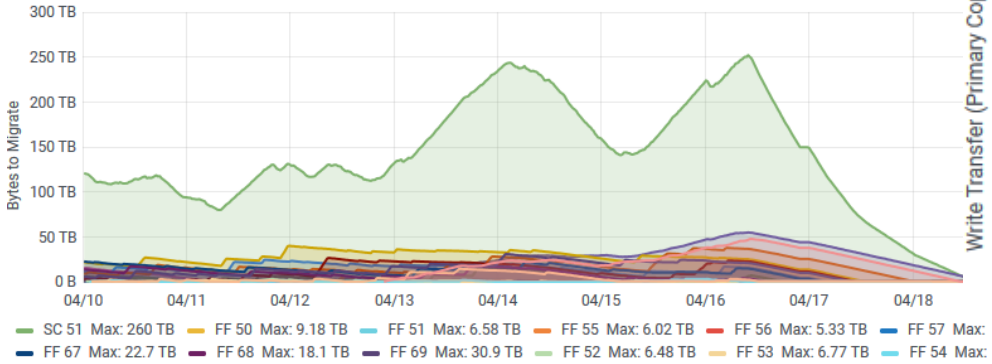
- use only one drive per file family to write data on tape
- use read-optimized aggregates
- aggregate as many files as possible
  - Max bytes in aggregate 300GB
  - Max files in aggregate 5000
- COS Migration order Directory

Monitoring

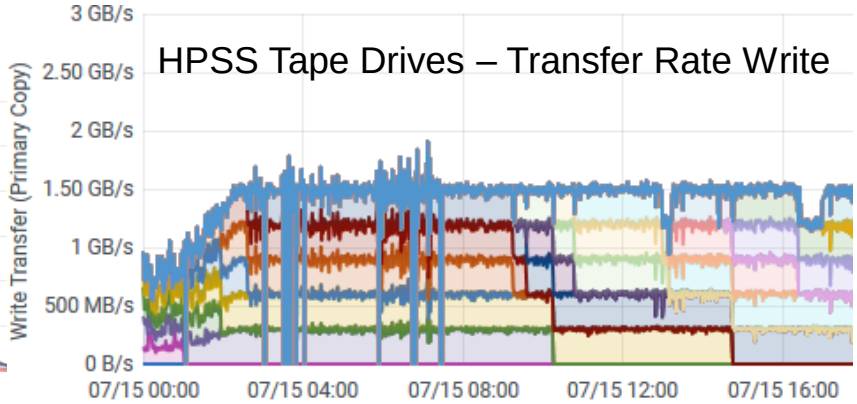
- HPSS cache
  - total # files and space used
  - purable # files and space
  - # files and space used to be migrated
  - # files and space staged
  - I/O rates
- Tape Drive rates
- Cartridge transfer rates
- # drives used in parallel by experiment
- Visualisation using Grafana
- rsyslog sends data to LogStash/ELK/Kibana

HPSS Monitoring

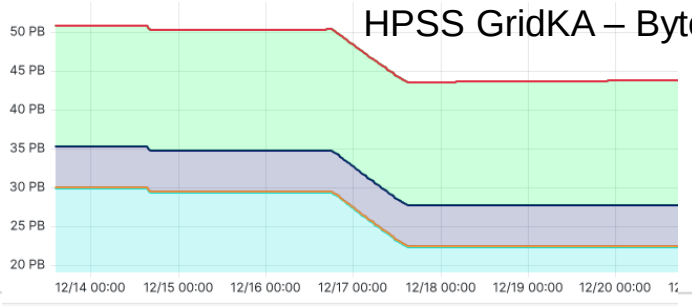
HPSS GridKa Cache – bytes to migrate



HPSS Tape Drives – Transfer Rate Write

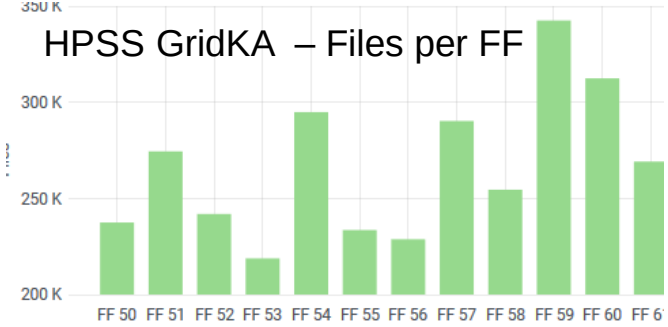


HPSS GridKA – Bytes per CoS



	Last
COS 51	26.7 PB
COS 61	123 TB
COS 71	5.55 PB
COS 81	31.6 PB
Total	63.9 PB

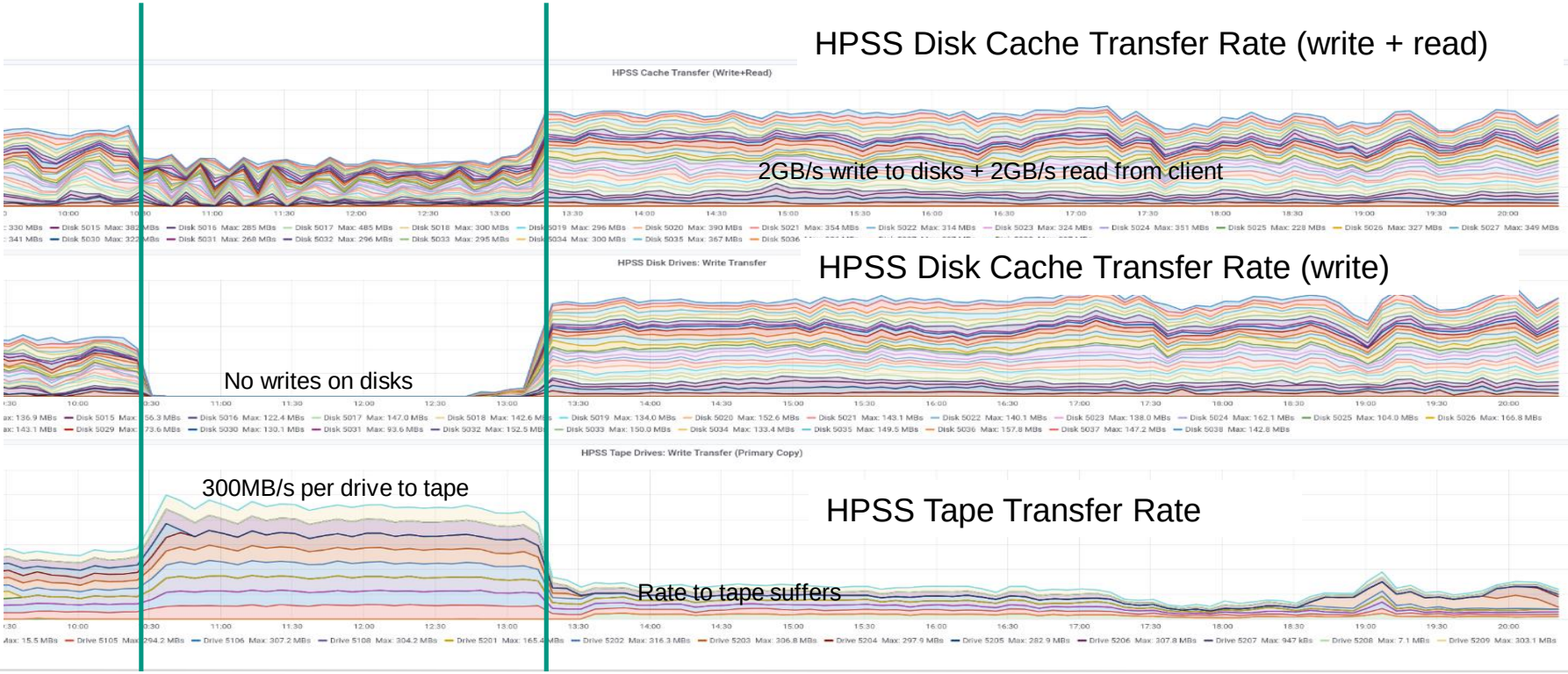
HPSS GridKA – Files per FF



Migrate GridKa data from TSM to HPSS issues

- HPSS Disk Cache throughput issue
  - NetApp E5700 120 SAS HDDs
    - tested throughput outside HPSS 12GB/s (50% read, 50% write)
  - HPSS cannot get more than 6 GB/s
  - Tried different configurations
    - 1 DDP with 26 volumes
    - 6 DDPs with 2 volumes each
  - Migration to tape is adversely affected
  - We had to reduce maximum number of tape drives for migration in order to get better data rate per tape drive
- Replacement of spinning disks by SSDs increased the drive rate to its maximal rate

Migrate GridKa data from TSM to HPSS issues



HPSS Recall Tests - Client

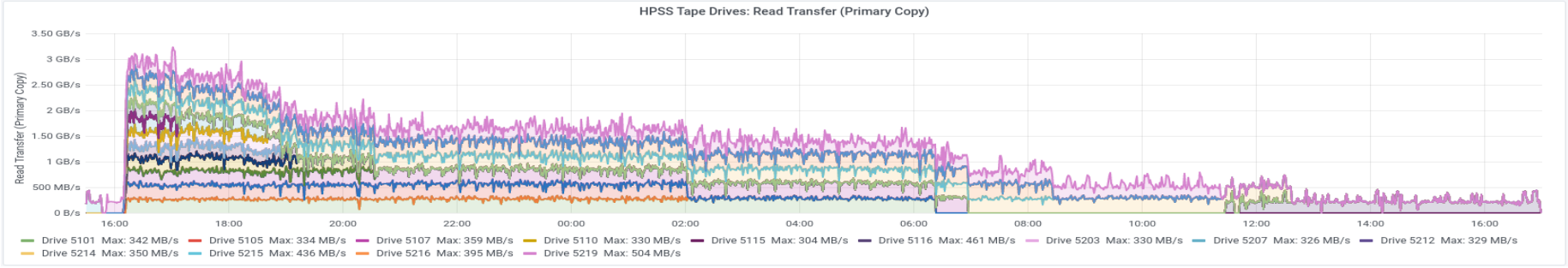
- HPSS FUSE
- Quaid
- PFTP:
  - proven client
  - one more component involved: pftp_server
- HPSS API:
  - Endit

dCache HPSS Interface

- rewrite ENDIT dCache Interface to TSM to access HPSS
  - started implementation for reading from HPSS
    - to be able to read the migrated files
  - the interface will wait
    - either for a number of files to be read from hpss
    - or for a specific time interval
    - configurable parameters
  - will sort the files on cartridge basis
  - will sort the files on aggregate basis
  - will start a staging process per aggregate using FAR
    - only one file from an aggregate

HPSS Recall Tests – FAR

- Successfully read 42314 files from 10 tapes in less than 24 hours with 10 drives available.
- The files were in 474 aggregates so there were never more than that number read requests in HPSS at the same time.
- Aggregates not staged in linear tape order; looks like TOR does use RAO info and schedules whole aggregates as a unit.



GridKa Transfer rates

- Max write rate (disk -> tape): ~2.0 GB/s (~390 MB/s per tape drive)
  - 8 tape drives used
- Max read rate (tape -> disk): ~4.0 GB/s (~380 MB/s per tape drive)
  - 14 tape drives used


Thank you for your attention
Contact: dataprotection@lists.kit.edu