Karlsruhe Institute of Technology

Storage at GridKa
A technical overview and outlook

The current situation of storage at GridKa and upcoming challenges.

Steinbuch Centre for Computing (SCC)

KIT – University of the State of Baden-Wuerttemberg and
National Research Center of the Helmholtz Association

Outlook

- Introducing the staff
- Overview
  - dCache Instances
  - xrootd
  - Tape backend
- Detailed setup
  - Fileserver
- Monitoring
  - Availability and Performance
- Upcoming challenges
  - in GridKa
  - in KIT (LSDF)

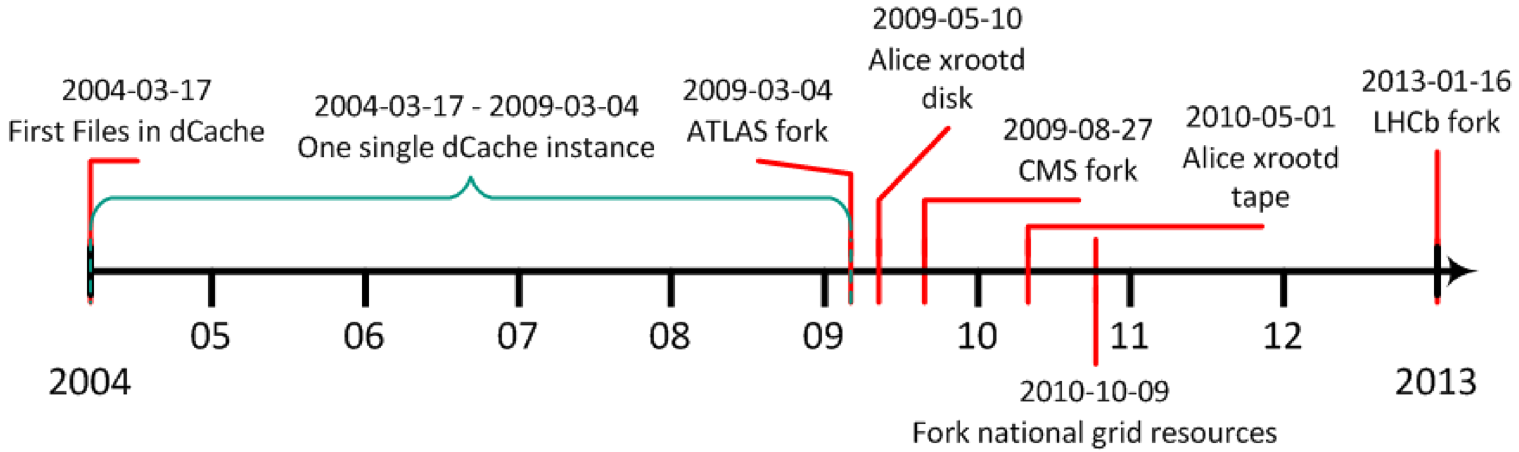
Introducing the staff

- 8 people for running all storage at GridKa until 2013.
  - But only 5.5 FTE from now on.
  - Currently many job opportunities at GridKa, so please tell anybody you know who might be interested.

<table>
<tr><th rowspan="2">Storage (Hardware)</th><th colspan="3">Storage (Software)</th><th rowspan="2">Tape Connection</th></tr>
<tr><th>GPFS</th><th>dCache</th><th>xrootd</th></tr>
<tr><td colspan="2">Jolanta Bubeliene</td><td></td><td></td><td>Martin Beitzinger</td></tr>
<tr><td colspan="3">Stephanie Böhringer</td><td></td><td>Dorin-Daniel Lobontu</td></tr>
<tr><td></td><td></td><td colspan="2">Verena Geisselmann</td><td></td></tr>
<tr><td></td><td></td><td>Xavier Mol</td><td></td><td></td></tr>
<tr><td></td><td></td><td colspan="2">Christoph-Erdmann Pfeiler</td><td></td></tr>
<tr><td></td><td></td><td>Doris Ressmann</td><td></td><td></td></tr>
</table>

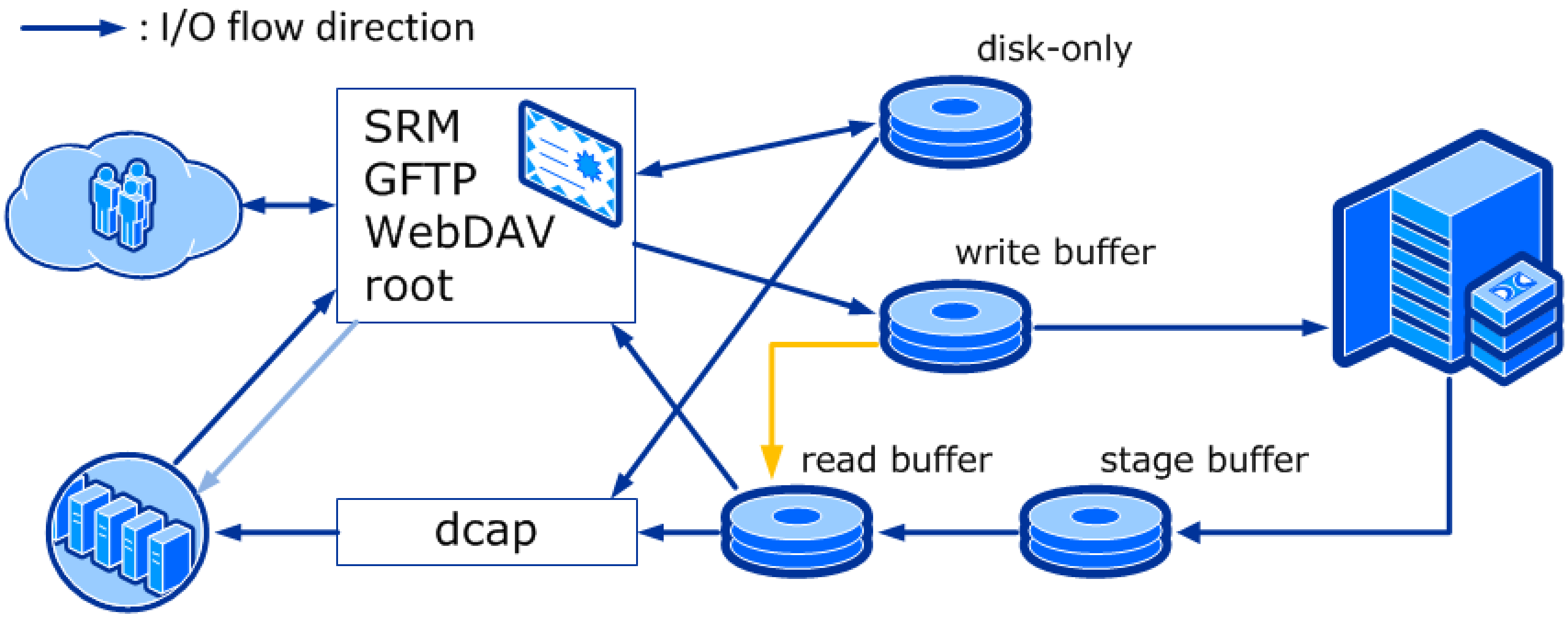
Overview – SE Instances

- Started with one big dCache setup for all experiments.
- Experiment’s demands are in conflict.
  - ATLAS emphasis on disk-only usage and SpaceTokens
  - CMS - quite the opposite - using dCache as tape buffer
  - LHCb focus on SpaceTokens, requirements substantially lower
  - Alice focused on xrootd
- All users suffer from problems caused by individuals.



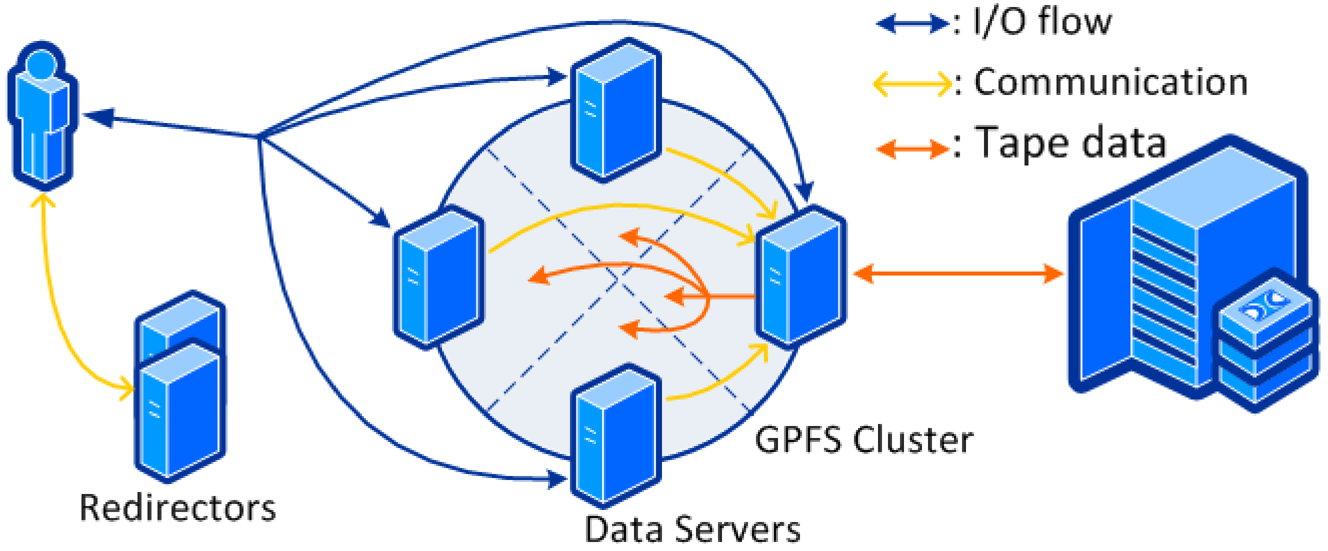
Overview – Typical dCache Setup

- The same for all dCache instances.
  - Only exception: LHCb may read from write buffer.



Overview – xrootd

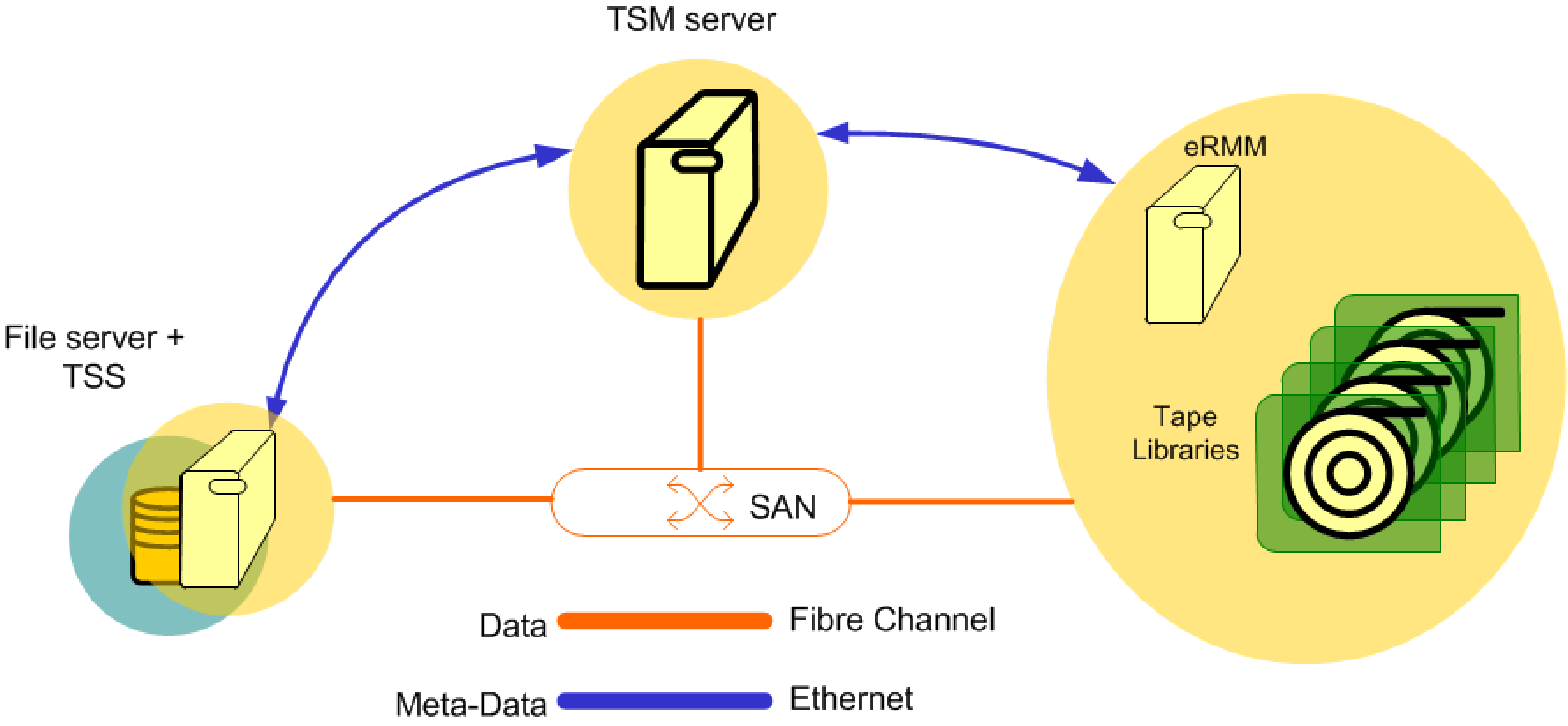
- Two independent SEs with xrootd: disk-only and tape buffer.
- xrootd tape buffer:



- Making every cluster node serve all data of the cluster.

Overview – Tape Connection

- Tape management software is Tivoli Storage Manager (TSM) combined with Enterprise Removable Media Manager (eRMM) as library virtualization software.
  - Only one tape library from the perspective of TSM.



Overview – Tape Connection

- Once, all file servers could write to and read from tape archive if necessary.
    - Required a lot of connections via SAN.
    - Chaotic data flow to/from tape.
- Limited number of nodes reading from tape for a SE (stage buffer).
    - Allows for better optimization of request processing.
    - Writing for every server possible.
- Maybe also bundle writing through few machines for dCache in the future.
    - Better optimization for archiving jobs.
    - Less administration overhead.

Overview – SEs in Numbers

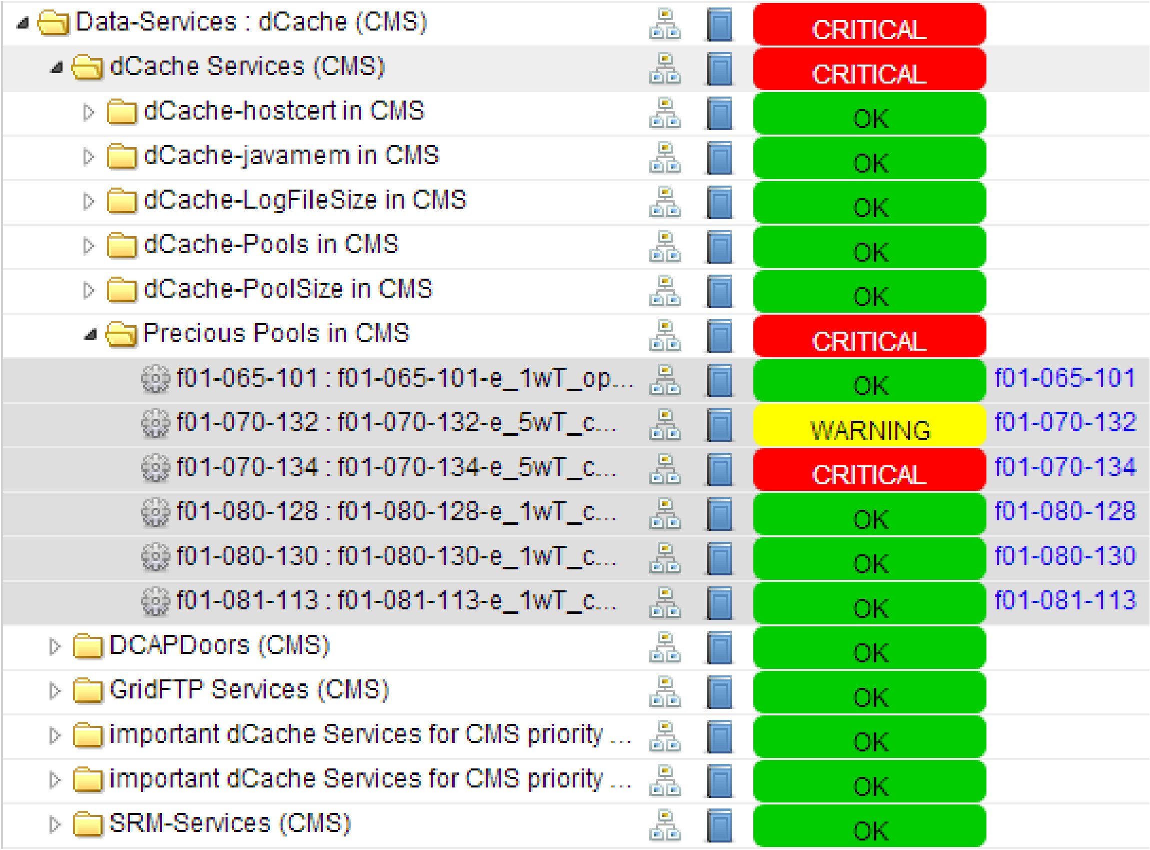
data volume always [TB]	Alice Xrootd	ATLAS dCache	CMS dCache	LHCb dCache	Total
Admin nodes / data servers	6 / 14	9 / 23	9 / 24	9 / 14	33 / 81
Disk space available (disk-only / tape buffer)	2.060 / 640	3.351 / 185	191 / 1.981	1.408 / 187	7.010 / 2.993
Data stored on disk	2.419	3.338	2.045	1.338	9140
Tape volume pledged	5.520	4.500	5.700	1.054	16.774
Data stored on tape	2.540	3.932	3.496	1.631	11.599
Data transferred 2012 (in)	3.100	405	332	548	4.385
Data transferred 2012 (out)	22.000	35.743	34.627	14.407	106.776
Volume archived on tape 2012	-	1.505	2.221	1.295	5021
Volume staged from tape 2012	-	10.969	5.177	2.885	19.031

Setup – File Servers

- File servers are machines with lots of RAM and medium number of cores.
  - Preferred 32-64 GB RAM
  - Modern machines 16+ cores
- Disk space is provided by DDN RAID-6 GPFS file systems.
  - Every two file servers form a GPFS cluster.
- Storage connected via SAN to the file servers.
- All file servers have 10 GE interface.
- Deployment of file servers with ROCKS or in-house tool (“CluClo”).
- Configuration management tool in validation phase.
  - Puppet or CFEngine 3
  - Currently CFEngine 2

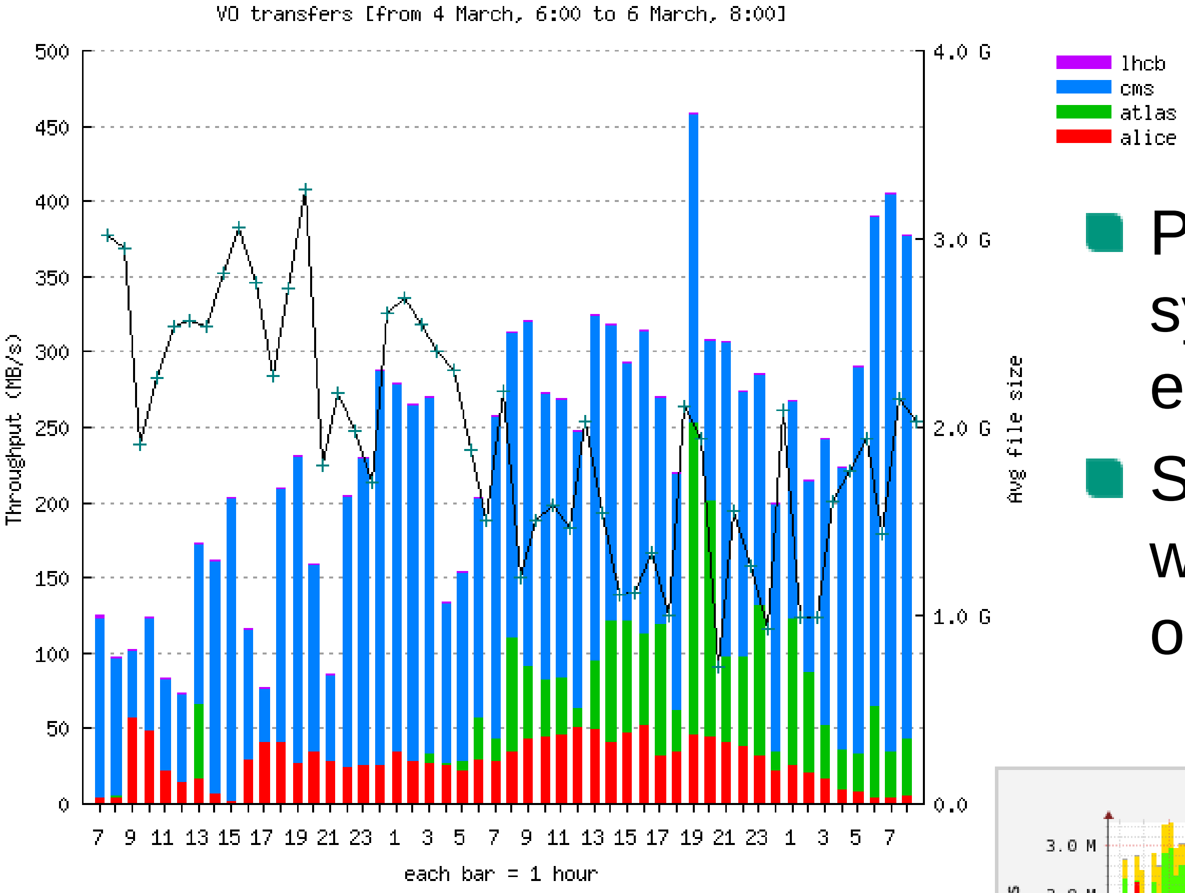
Monitoring – Availability

- Switched from Nagios to Icinga in April 2012.
- Hierarchical organization of monitored services
- On-call service steered by Icinga.
- Certain number of critical services will trigger state change, which then triggers oncall alarm.

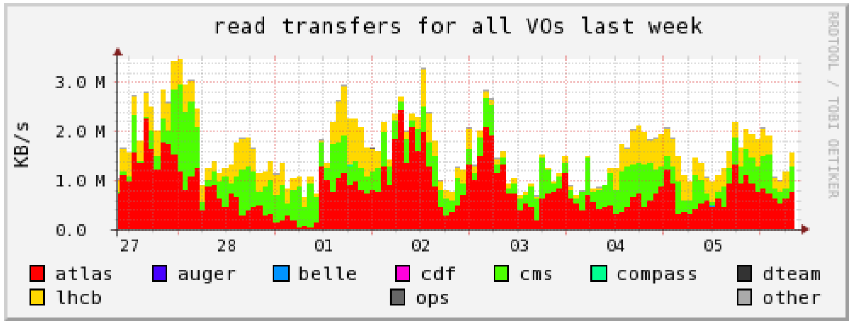


◢ Data-Services : dCache (CMS)	CRITICAL	
◢ dCache Services (CMS)	CRITICAL	
▷ dCache-hostcert in CMS	OK	
▷ dCache-javamem in CMS	OK	
▷ dCache-LogFileSize in CMS	OK	
▷ dCache-Pools in CMS	OK	
▷ dCache-PoolSize in CMS	OK	
◢ Precious Pools in CMS	CRITICAL	
f01-065-101 : f01-065-101-e_1wT_op...	OK	f01-065-101
f01-070-132 : f01-070-132-e_5wT_c...	WARNING	f01-070-132
f01-070-134 : f01-070-134-e_5wT_c...	CRITICAL	f01-070-134
f01-080-128 : f01-080-128-e_1wT_c...	OK	f01-080-128
f01-080-130 : f01-080-130-e_1wT_c...	OK	f01-080-130
f01-081-113 : f01-081-113-e_1wT_c...	OK	f01-081-113
▷ DCAPDoors (CMS)	OK	
▷ GridFTP Services (CMS)	OK	
▷ important dCache Services for CMS priority ...	OK	
▷ important dCache Services for CMS priority ...	OK	
▷ SRM-Services (CMS)	OK	

Monitoring – Performance



- Performance mostly synonym to transfer efficiency.
- Statistics gathered with Ganglia and other tools.



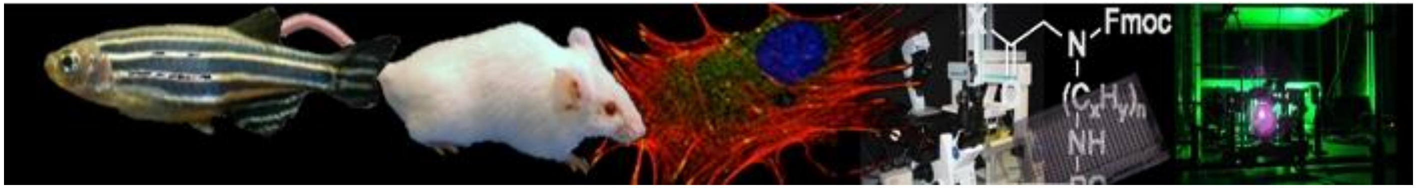
Upcoming Challenges in GridKa

- Update of dCache Golden Release (all dCache SEs)
  - Change of installation and configuration directories (FHS-compliant)
  - Change of authentication and authorization model (gPlazma v1 → gPlazma v2)
- New protocols for ATLAS
  - http/s with WebDAV
  - Joining Federated ATLAS Xrootd (FAX)
- Paradigm shift for CMS
  - Less tape buffer, lots of disk-only space
    - Explicit control over which files when get flushed to tape
- Improve tape throughput significantly.

GridKa's Younger Sister – Large Scale Data Facility

LSDF

KIT Karlsruhe Institute of Technology



- LSDF is about Data Management, Data Analysis and the Data Life Cycle
- Support for data intensive computing for in principle all sciences
  - Biology, materials research, climate research, geology, …
  - Institutes of KIT and the state of Baden-Württemberg
  - Cooperating with EU Projects EUDAT, DARIAH

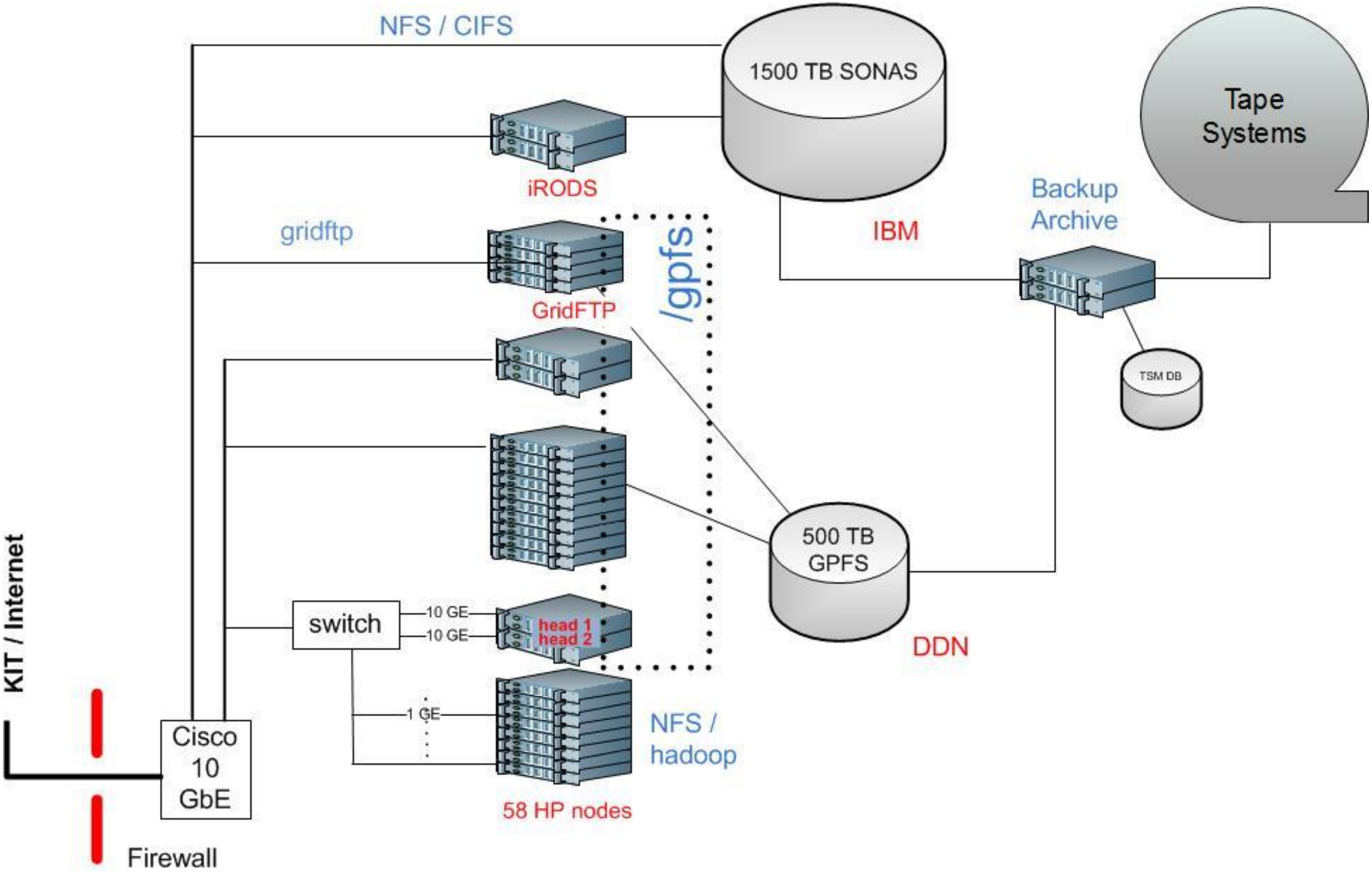
EUDAT
European Data Infrastructure

DARIAH-DE
Digital Research Infrastructure
for the Arts and Humanities

GridKa’s Younger Sister – Large Scale Data Facility

- Heavy LSDF users (> 1 PB)
  - Synchrotron light source ANKA: Tomography and topology beamline
  - Single Plane Illumination Microscopy (aka Lightsheet Microscopy)
    - Generating 1000s of high-res images per day
- Leveraging experience from GridKa (though without ‘Grid’)
  - Consolidation of GridKa and LSDF expertise and resources in the future

LSDF – Hardware and Network Layout



Finally…

- Thank you for your attention!
- Any questions?