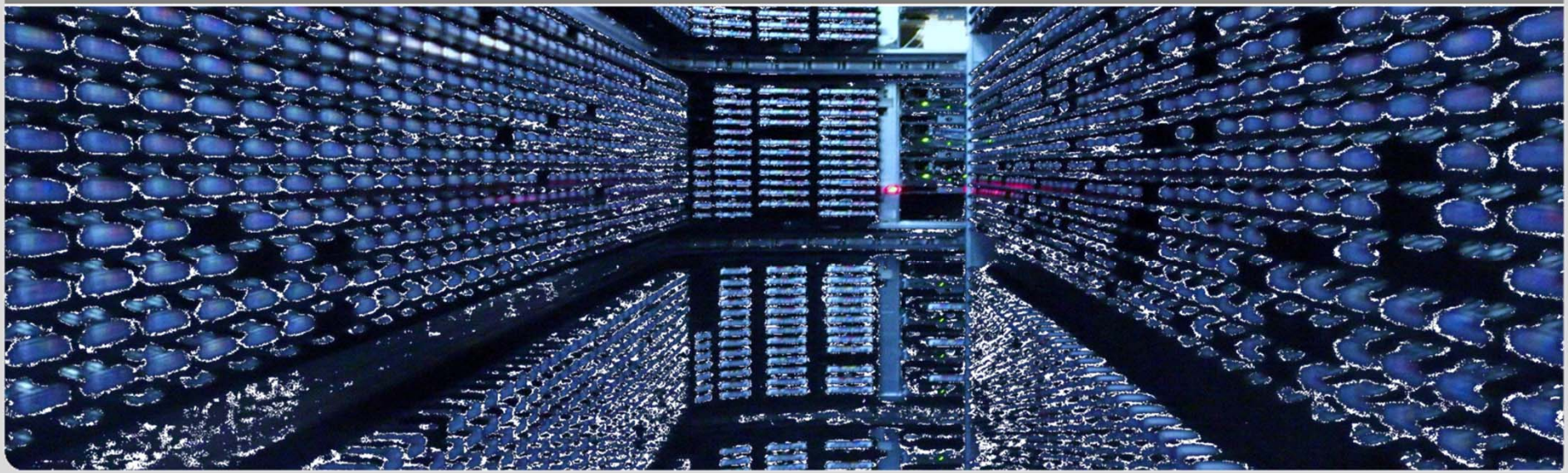
Karlsruhe Institute of Technology

SCC – mission and vision

Visit at HSS lab

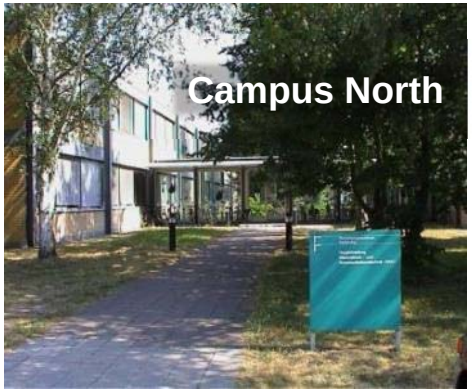
STEINBUCH CENTRE FOR COMPUTING - SCC



Steinbuch Centre for Computing

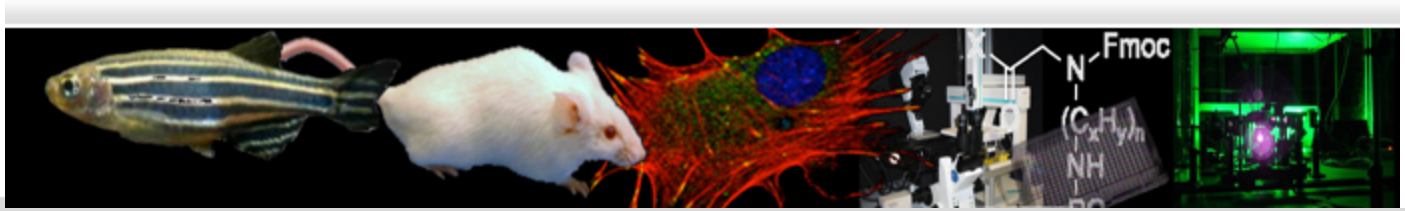
Computer Centre of the Karlsruhe Institute of Technology

- **IT for the University Campus**
  - 6000 staff , 20000 students
- **Research**
- Data Management, Data Analysis and secure IT Federations
  - Large Scale Data Management & Analysis
  - GridKa
  - Federated Identity Management
- Computational Science and Engineering
  - High Performance Computing
  - Simulation Laboratories
- Dynamic IT Infrastructures
  - Cloud Computing
  - Virtualisation
  - Web Engineering
  - IT and Service Management
- **Operation of large facilities**



Large Scale Data

- **GridKa – LHC Tier 1 centre since 2002**
  - Supporting all 4 LHC experiments
  - Currently 14 PB storage, 16000 cores, 10 Gb/s networking
  - Operations and software dedicated to physics off-line computing
  - Using gLite grid middleware
- **Large Scale Data Facility (LSDF) since 2009**
  - Support for data intensive computing for (in principle) all sciences
  - Biology, Materials research, Climate research
    - Storage for Baden-Wuerttemberg
  - 4.5 PB storage, 500 cores computing, 10 – 100 Gb/s networking
- **HPC for KIT and Baden-Wuerttemberg**
  - 3 HPC Systems ranging from 30 to 135 Tflops, 300 TB Storage



Fact sheet

- GridKa
  - SA9900, SFA10K
  - GPFS in multiple clusters (500 – 1000 TB)
  - xrootd and dCache, CVMFS, for storage management
- LSDF
  - SFA12K, DS5300
  - GPFS and SONAS
  - Industry standard protocols
- HPC
  - SA9900 HPC Online storage
  - Shared FS with 2 clusters
  - Lustre

tape env

Campus South

Grau XL 5300 slots, LTO3, LTO4

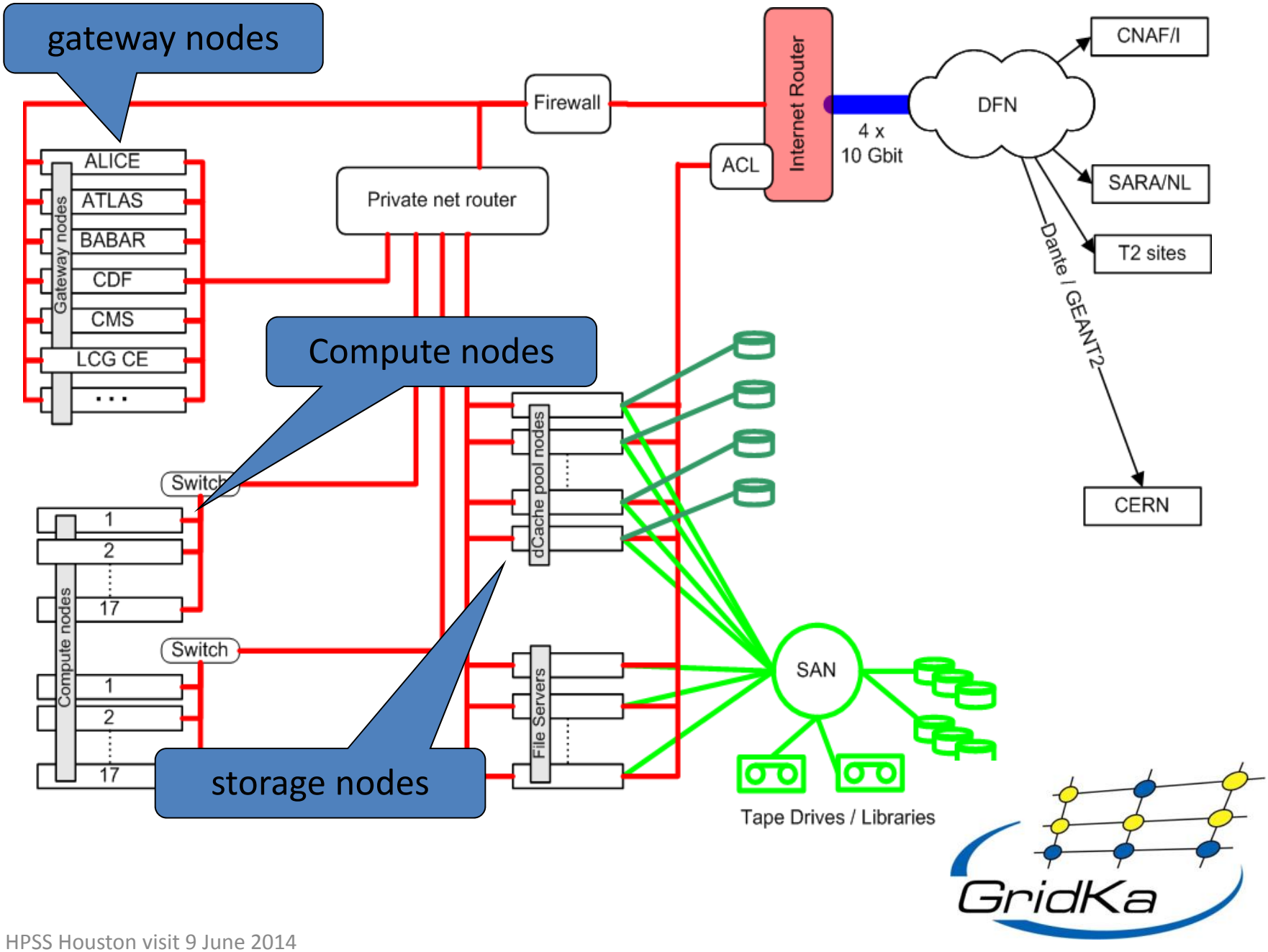
STK 01, 10000 slots 16 LTO4, 12 LTO5

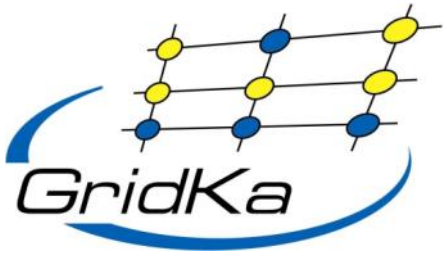
STK 02, 10000 slots, 10 LTO5, 2 T10Kd

KIT Backup

8400 slots, 22 LTO5

6000, 24 LTO4

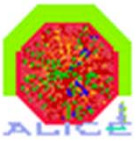

gateway nodes
Firewall
Internet Router
ACL
4 x
10 Gbit
DFN
CNAF/I
SARA/NL
T2 sites
Dante / GEANT2
CERN
Private net router
ALICE
ATLAS
BABAR
CDF
CMS
LCG CE
...
Gateway nodes
Compute nodes
Switch
1
2
17
Compute nodes
Switch
1
2
17
dCache pool nodes
File Servers
SAN
Tape Drives / Libraries
storage nodes
GridKa



Karlsruhe Institute of Technology

Jobs running	1298
Jobs queued	5
LCG Jobs running	1298
LCG Jobs queued	4

Jobs running	72
Jobs queued	0
LCG Jobs running	0
LCG Jobs queued	0



Jobs running	9
Jobs queued	2
LCG Jobs running	9
LCG Jobs queued	2

Jobs running	2
Jobs queued	0
LCG Jobs running	2
LCG Jobs queued	0

Jobs running	49
Jobs queued	0
LCG Jobs running	49
LCG Jobs queued	0

Jobs running	1
Jobs queued	0
LCG Jobs running	0
LCG Jobs queued	0

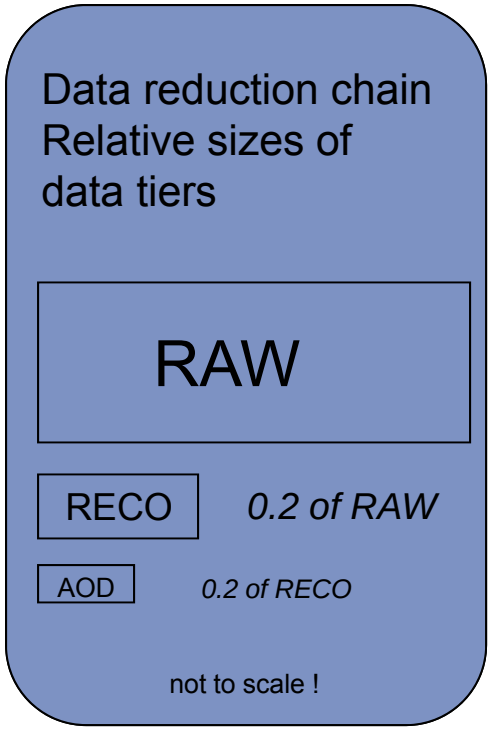
Jobs running	10
Jobs queued	0
LCG Jobs running	10
LCG Jobs queued	0

Jobs running	1
Jobs queued	0
LCG Jobs running	0
LCG Jobs queued	0

LSDF

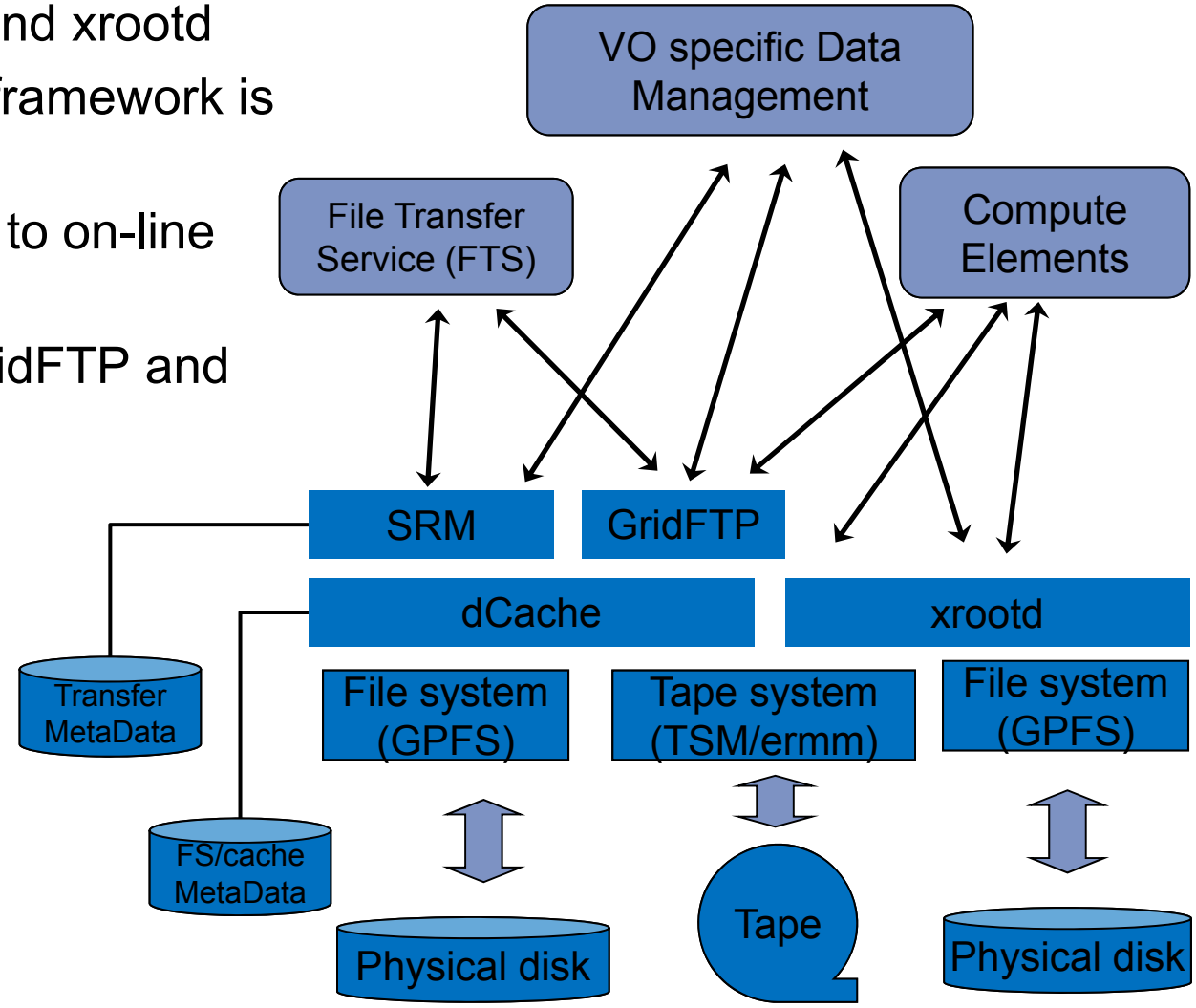
Data (reduction) in HEP

- Detector data
  - RAW - tracks, hits : size ~2.5MB
    - needs reconstruction
  - RECO - detailed reconstructed info, particles
    - suitable for detector studies and reconstruction code development
  - AOD - most used objects for analysis
    - (**A**nalysis **O**bject **D**ata)
    - this is what end scientists use to run their jobs.
  - Skims - selected channels for focused research groups and individual scientists
    - they are just pointers to AOD objects
  - Ntuples - suitable for download to laptop and to run ntuple analysis (apply selections, cuts)
- Calibration and condition data
  - must be available at every job
  - side band data flow using distributed databases



Storage in WLCG / GridKa

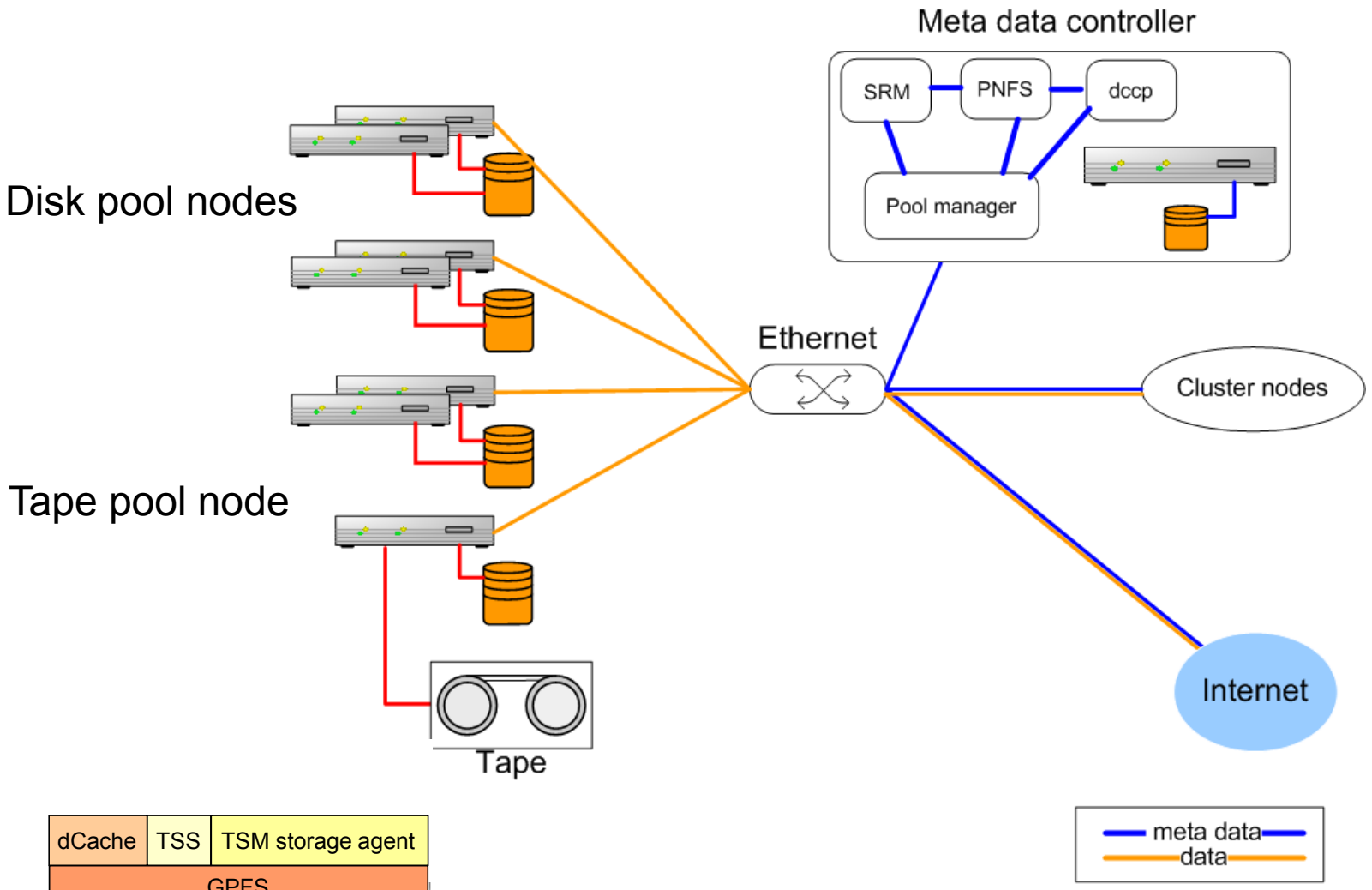
- Based on dCache and xrootd
- Data management framework is VO specific
- SRM based access to on-line and archive storage
- WAN access via GridFTP and managed by FTS

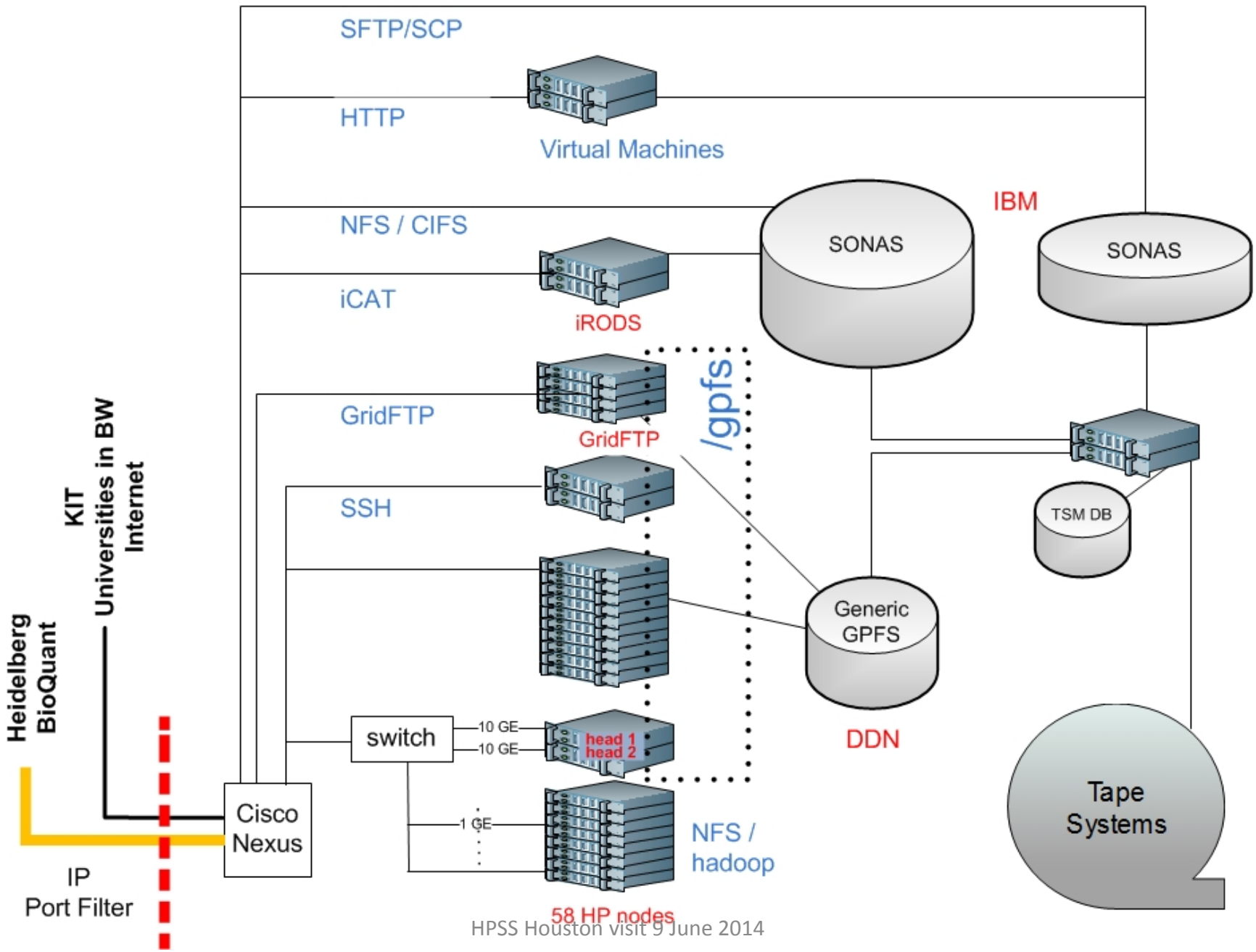


What is dCache?

- Think of it as file system (with some restrictions)
- Unified name space
- Disk pool management: unifies several storage units into one
- Data is distributed on a number of servers with disk pools
- Load balancing and hot spot detection
- Support for HSM backend
- Development DESY(Deutsches Elektronen-Synchrotron) and FNAL(Fermi National Accelerator Laboratory)

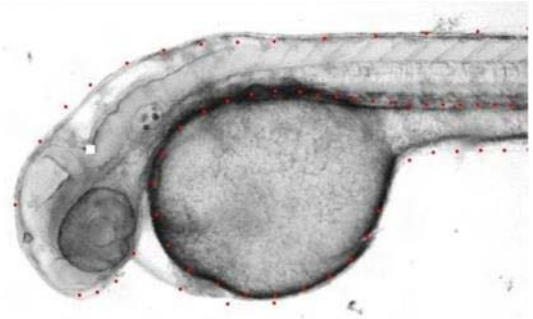
dCache architecture

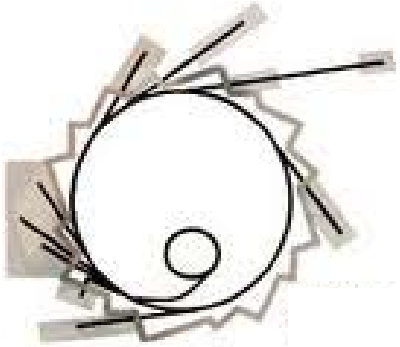


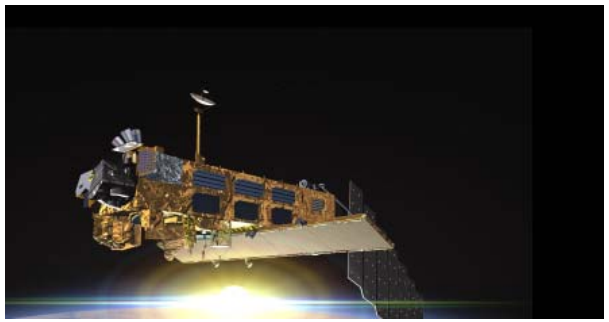

SFTP/SCP
HTTP
Virtual Machines
NFS / CIFS
iCAT
iRODS
IBM
SONAS
SONAS
GridFTP
GridFTP
/gpfs
SSH
TSM DB
Generic
GPFS
DDN
Heidelberg
BioQuant
KIT
Universities in BW
Internet
switch
10 GE
10 GE
head 1
head 2
Cisco
Nexus
1 GE
NFS /
hadoop
58 HP nodes
Tape
Systems
IP
Port Filter

LSDF Data

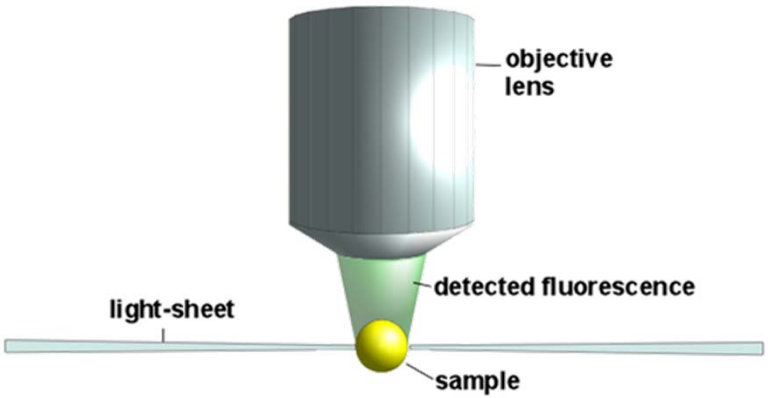
- Biology
  - High Throughput Microscopy 1000 TB/year
  - Gene Sequencing 50 TB/year
  - PByte archives for partner institutes
- Synchrotron radiation (ANKA)
  - Various experiments
  - 200 TB/year (2013) - 1000 TB/year (2014)
- Climate research
  - Various experiments (rate stays flat)
  - 150 TB/year (archive until 2026!!)

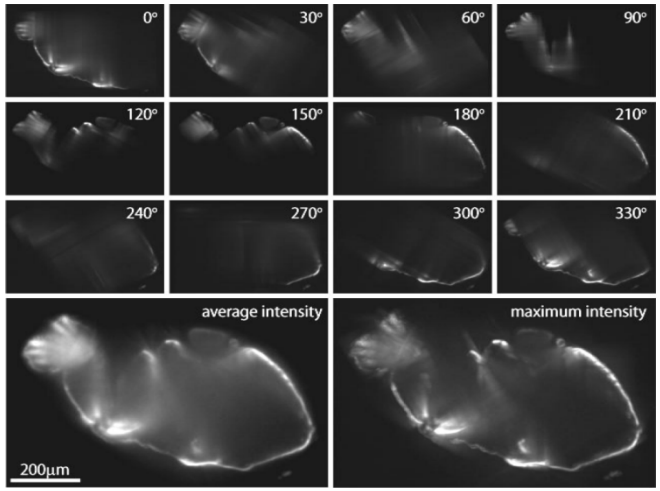






Application: Light Sheet Microscopy





Source: Uroš Kržič,
Multiple-view microscopy with light-sheet based fluorescence microscope, Dissertation, Heidelberg, 2009

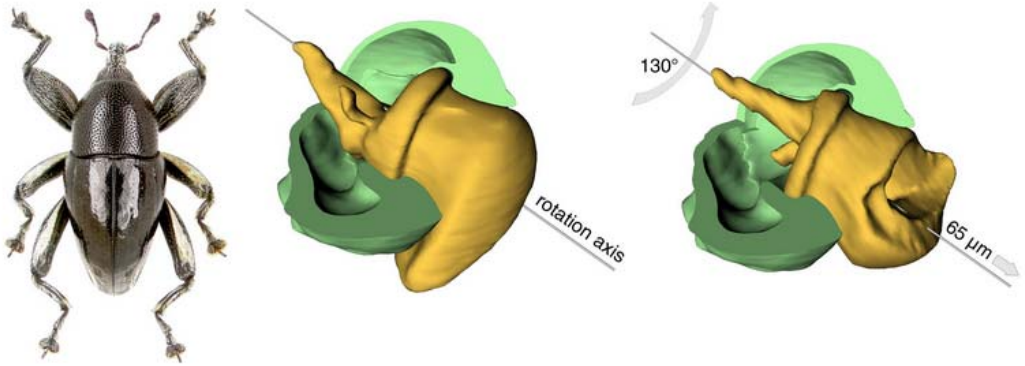
Novel microscope to observe **zebra fish embryos in vivo**:

- Very high 3D resolution
- Short data acquisition time

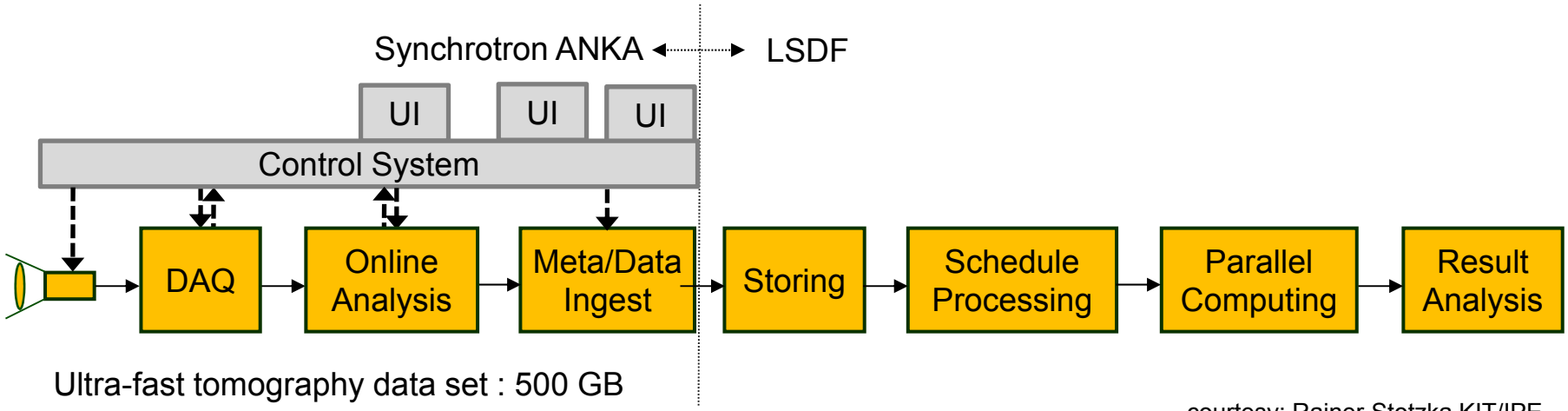
Each day one observation of the development of an embryo:

- Data: 7 TB/10 h → expected ~ 300 TB
- Challenges:
  - Meta data extraction
  - **Data transfer** source to LSDF > 250 MB/s
  - **Data intensive analysis** < 24 h: registration, fusion, segmentation, tracking, statistics

Application: Ultra-Fast Tomography



- Spatio-temporal observations
- Interactive 3D reconstructions based on tomographic volumes
- Development of
  - data management components within the ANKA beam line and the LSDF,
  - online analysis based on GPGPU,
  - novel reconstruction algorithms



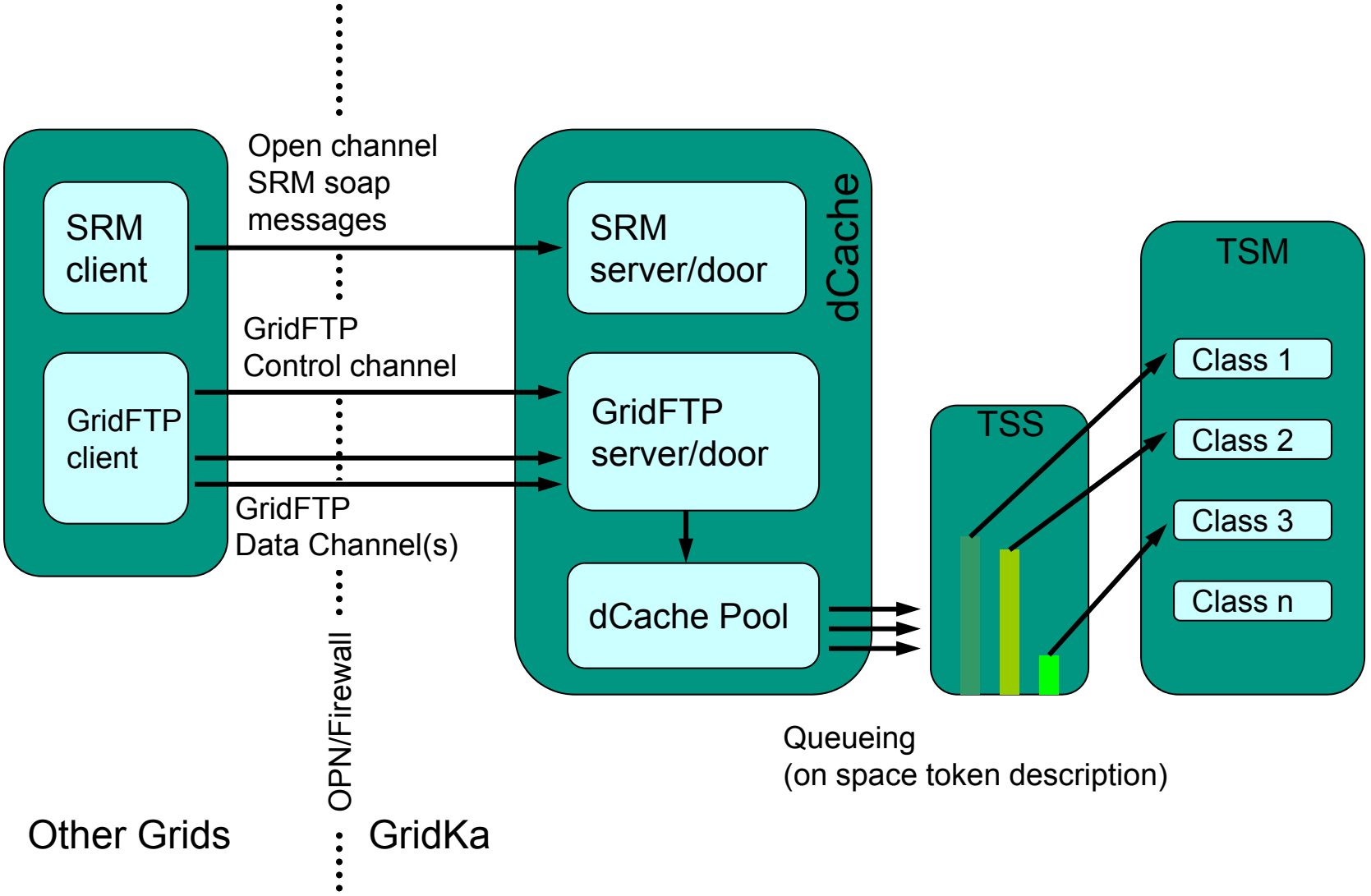
courtesy: Rainer Stotzka KIT/IPE

Connection to tape

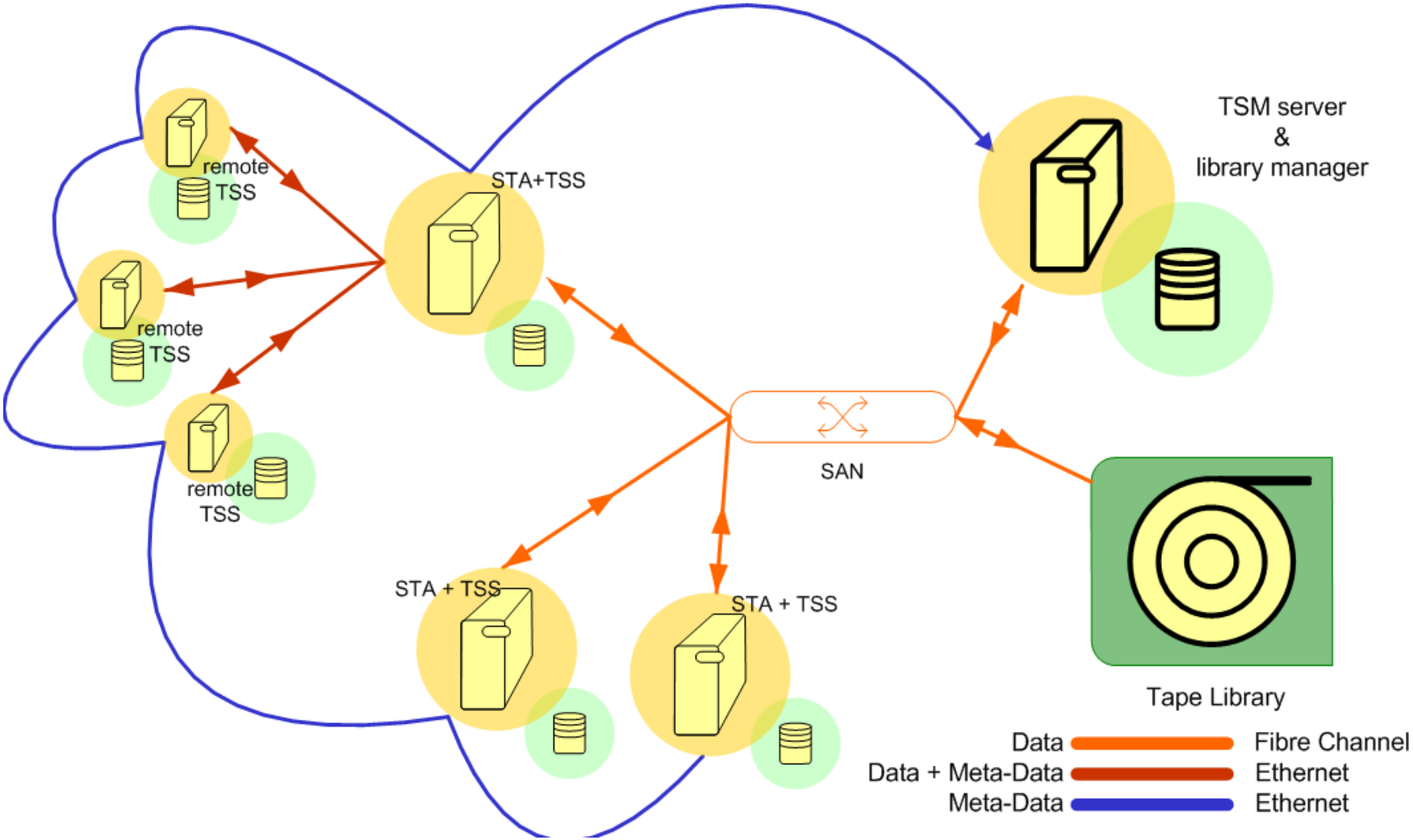
- backends needed for
  - dCache: GridKa
  - xrootd: GridKa
  - iRODS: EUDAT, HBP
- All these call backend scripts
  - uses get put del primitives
  - eg. dCache “put”

```
exec("$TSS",  "--migrate",  "--class=$CLASS",  "$desc", \ "$DSKFNAM",
"/$STORE/$GROUP/$PNFSID");
```

Data flow to the T1 - with dCache and SRM



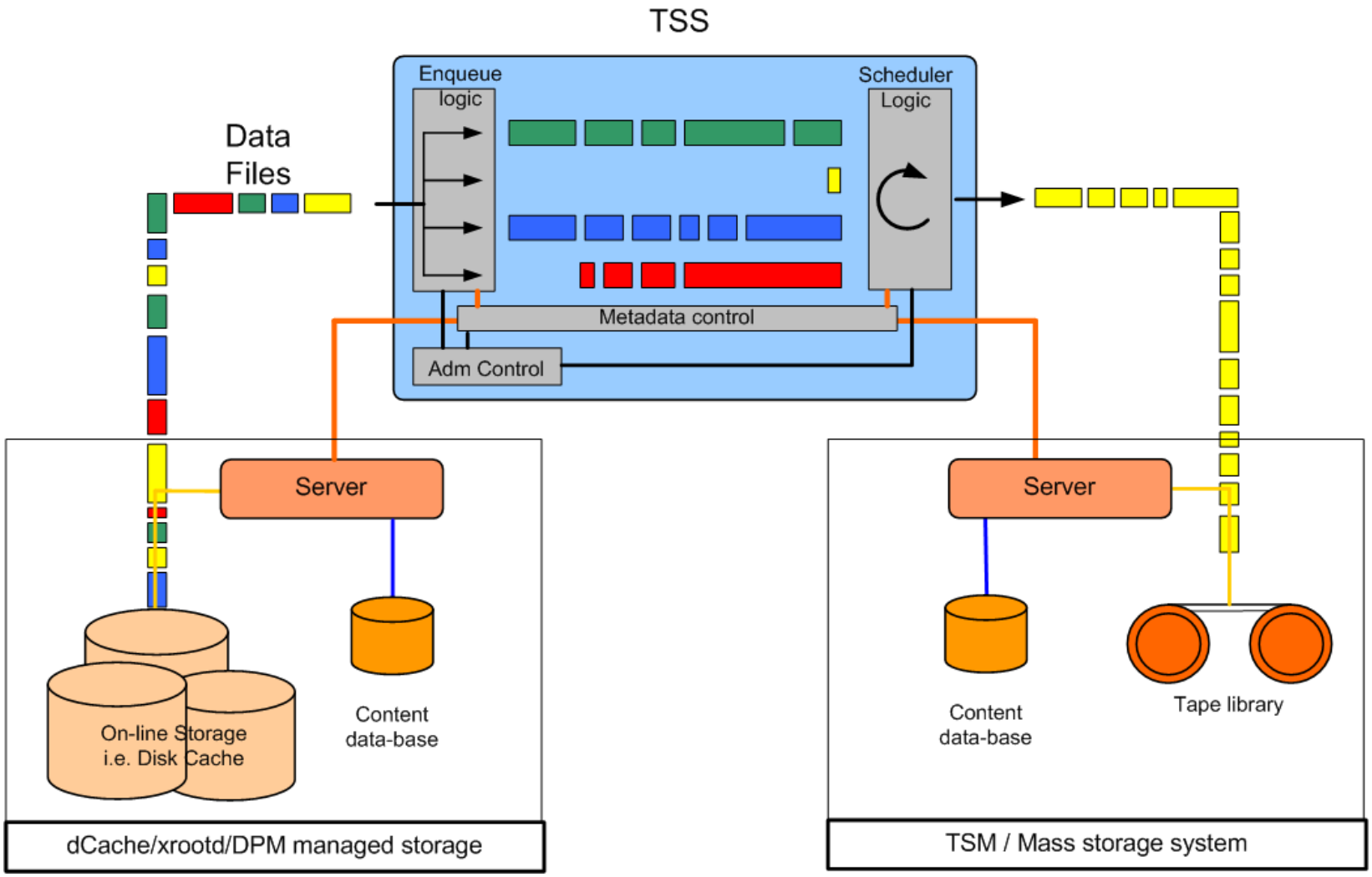
Archives with TSM



TSS

- Interface directly with TSM via the API
- Fan out for all dCache to tape activities
  - multiple operations: recall, migrate, rename, delete, query
- Runs on the TSM clients, storage agent or on the server proper
- Plug-in replacement for the TSM backend that comes with dCache
- Sends different type of data to different tape sets
- Queues recall requests on tape sequence order
- Allows to store an exact image of the logical global name space on tape
- command line interface to set running parameters, monitor the processing, run db queries (think of it as an alternative dsmc)

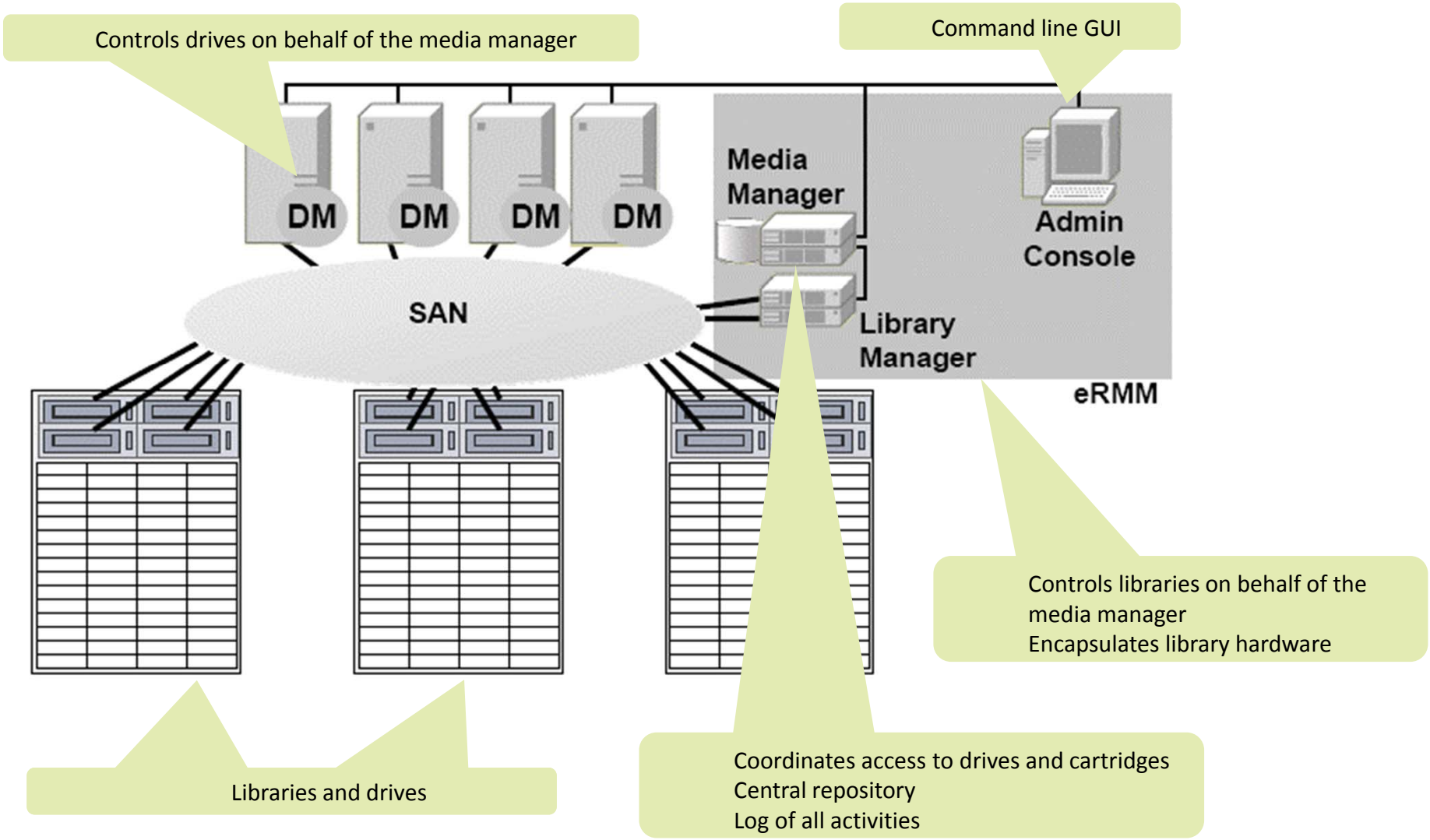
TSM staging server (TSS)



Enterprise removable media manager (eRRM)

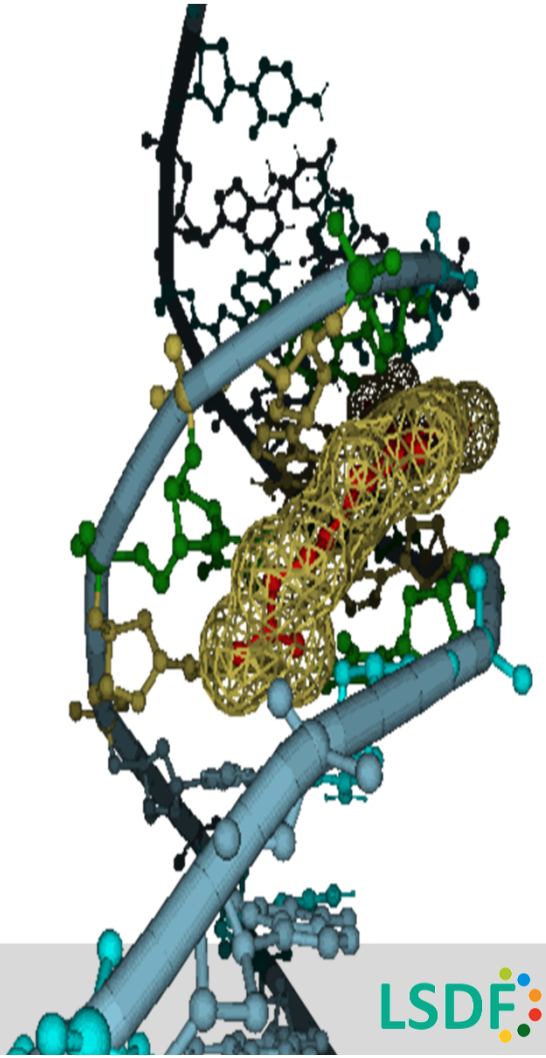
- Dynamic library and drive sharing across heterogeneous application boundaries and operating systems
- Dynamic drive and cartridge pooling
- Mount request queuing
- Policy based drive and cartridge allocation
- Policy based media life cycle management
- Integrated off-site media management and tracking (vaulting)
- Centralised access control, administration and reporting
- Advanced reporting and auditing

eRMM architecture

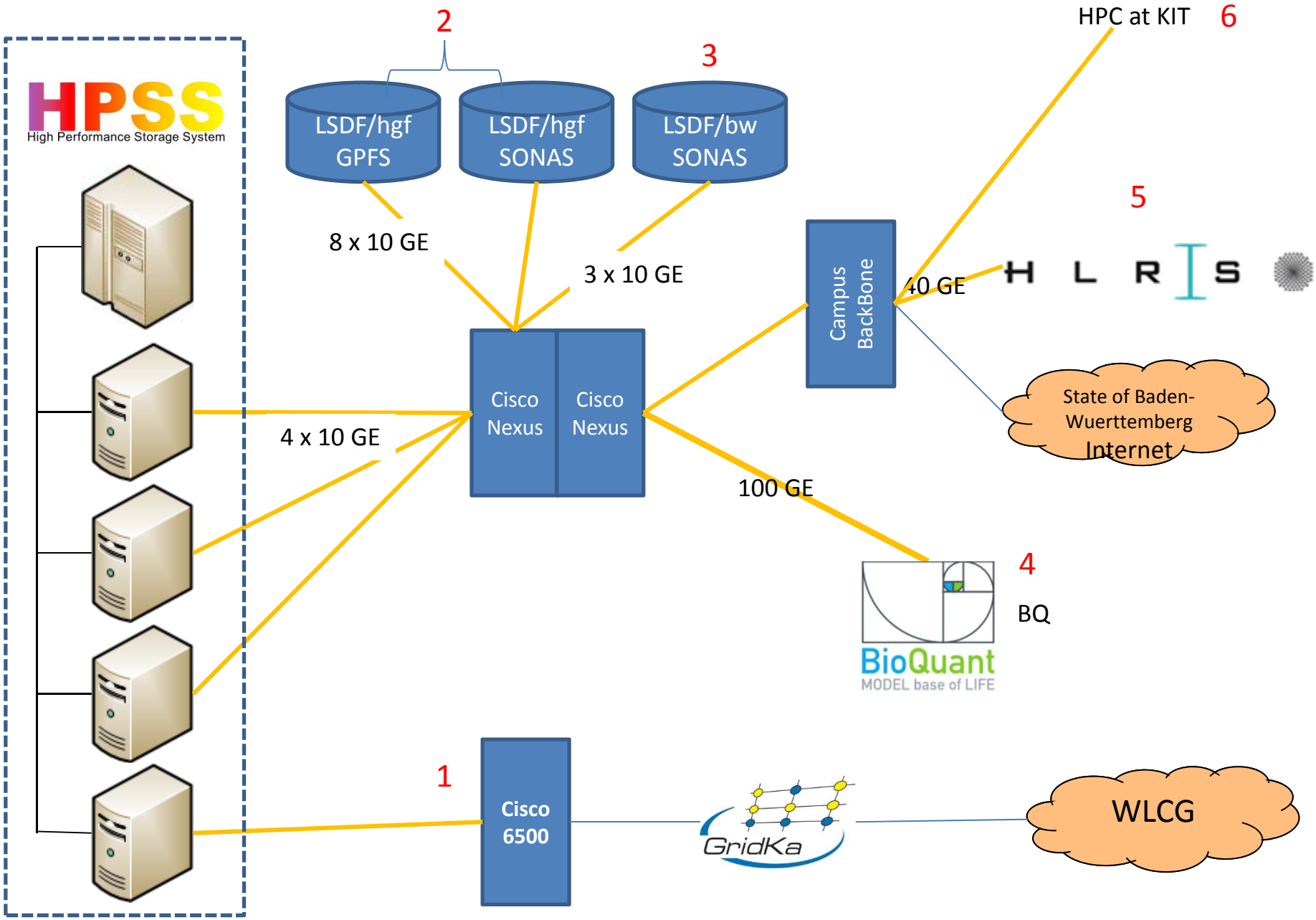


Other KIT customers (besides LSDF and GridKa)

- LSDF DIS
  - images, biology, materials research
  - small files, access via gftp, sftp, GHI/VFS
- BioQuant
  - images, genetics data
  - currently 4000 TB: backup via TSM
  - input as is (small files!!)
- HLRS
  - 1 - 2 PetaBytes yearly from 2015
  - users can and must tar data
- Libraries and State Archives
  - file upload via web tools
  - expect several hundreds of TB per year



Services


HPSS
High Performance Storage System
2
LSDF/hgf GPFS
LSDF/hgf SONAS
3
LSDF/bw SONAS
8 x 10 GE
3 x 10 GE
HPC at KIT
6
5
H L R S
Campus BackBone
40 GE
State of Baden-Wuerttemberg
Internet
Cisco Nexus
Cisco Nexus
4 x 10 GE
100 GE
4
BQ
BioQuant
MODEL base of LIFE
1
Cisco 6500
GridKa
WLCG

Update to Growth plans 2014 – 2016
Preliminary. Not valid until 1.8.2014

Project	Size PB	File Size Range	Growth PB/a	Client	Comment
HLRS	1	> 2 GB	1	scp	can use TAR
GridKa	12	> 2 GB	2	API	need queuing
LSDF-DIS (=HGF)	3	< 200 MB	2	gFTP	*)
LSDF-Heidelberg	4	< 200 MB	1	gFTP	*)
LSDF-EUDAT/HBP	1.5	< 200 MB	0.5	API	Object Storage
LSDF-State BW	0.5	< 200 MB	0.5	GHI	SONAS/NAS
KIT HPC	2	< 200 MB	0.5	API	Lustre HSM

*) a web based upload tool is being developed

HPSS Archive Project at KIT

- Project start is 1.1.2014
  - the HPSS project will not use the full installed capacity from the beginning. Instead project are added one after the other. Preparations for adding a KIT project to HPSS may run in parallel. Adding projects must be done the sooner the better. Hard deadlines are depicted below in the list of deliverables.
  - we will add hardware (movers, cache space) as needed.
  - Q3 and Q4 will be used to become familiar with the system, run tests, try out different clients and setups etc. KIT expects HPSS support when performing these trials. During Q4 the system will become production ready (KIT production readiness requires several formal steps to be taken)
- Deliverables (resource requirements are estimates that must be adapted 'in situ')
  - HPSS (production and test) ready for use: 1.7.2014
  - Accept and store data from HPSS at HLRS: 1.1.2015
    - the exact requirements will be determined the coming months.
    - KIT and HLRS will involve HPSS support for finding an optimal solution
    - 2 to 10 drives LTO5
  - Archives for BQ: 1.2.2015
    - 2 to 10 drives LTO
  - GHI on LSDF/hgf: 1.3.2015
    - 2 to 6 drives LTO
  - Migration from GridKa to HPSS start: 1.6.2015
    - 10 – 15 drives LTO for migration
    - 20 – 50 drives for production workload

HPSS discussion topics

- GHI
- Tape in the storage hierarchy adding storage class memory
  - combine object oriented storage to tape
- LTFS
  - can this format be used for tape archives and replace the HPSS format
- Large Media
  - how to effectively handle >> 100 TB media
- Queuing HPSS requests
  - reading tape is 30% of TSM
- Replacements (some of the) backups
  - LSDF-DIS has 3000 TB
- Replacing TSM in GridKa
  - connection from dCache and xrootd to TSM
  - 2 approaches to the solutions
    via API or GHI (I'm striking VFS because I do not think that is a viable solution).
  - infos from IN2P3 and BNL can serve as template
  - use FUSE?