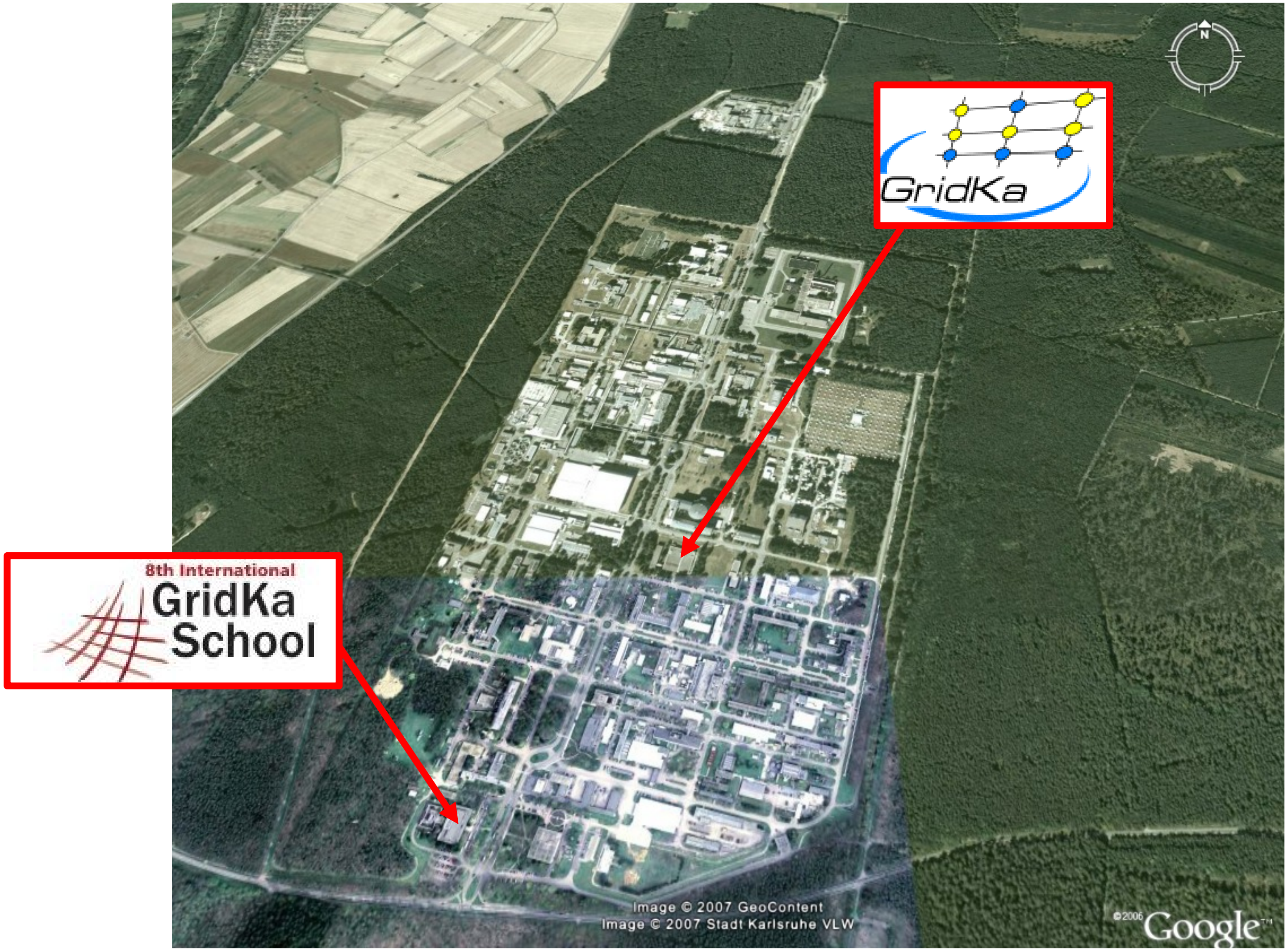
Karlsruher Institut für Technologie

The Tier-1 centre GridKa

Dr. Andreas Heiss

Steinbuch Centre for Computing

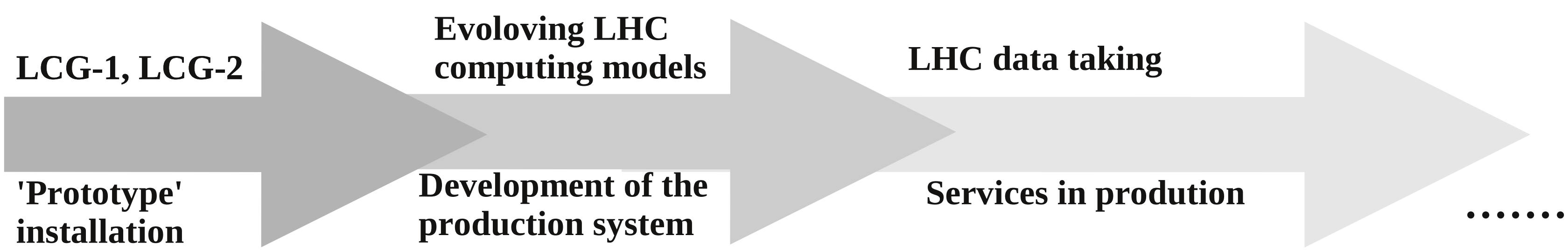
KIT north campus



Andreas Heiss

History

- 2001: Proposal of a „Regional Data and Computing Centre“ (RDCCG) by the Particle and Nuclear Physics Communities in Germany.
- 2002: Start of the project GridKa at (former) FZK
- Three project phases

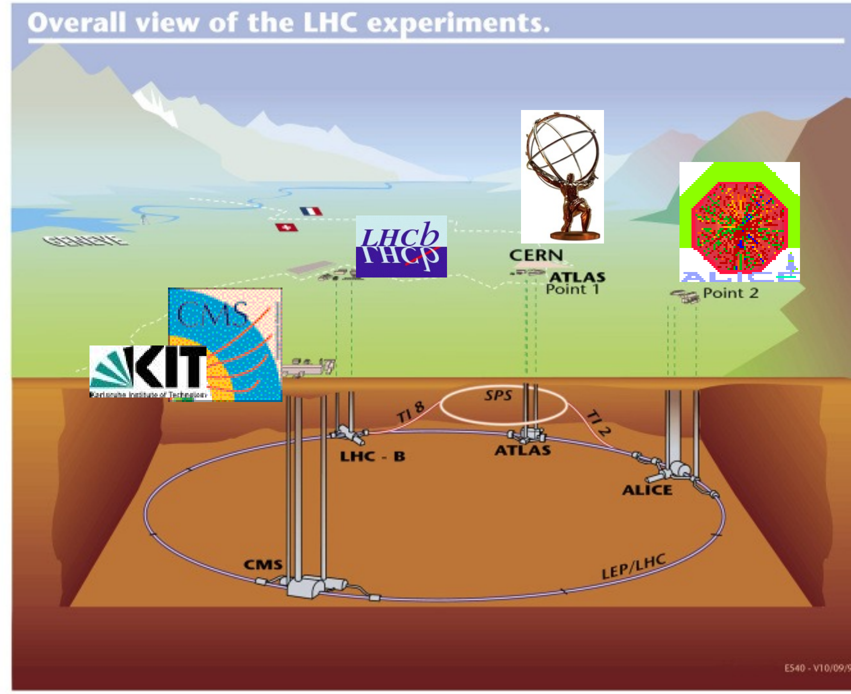


- Production site for non-LHC experiments (e.g. Tevatron: CDF, D0) long before the LHC start
  - gain experiences with HEP computing
  - test Grid techniques

- Phase 3 just started. Are we finished now?

GridKa today: resources and services for HEP and Astroparticle physics experiments

- GridKa supports all 4 of the big LHC experiments as a 'Tier-1' centre.
- GridKa is responsible for the storage and processing of approx. 14% of the total LHC data.
- GridKa supported non-LHC experiments:



- Resources for Compass, Babar, CDF, D0 remain approx. constant until end of data analyses.
- Grid test environment for Belle-II
- Resources for Auger

Projects

- GridKa paticipates in national and international projects and working groups:
  - Test setups
  - R&D
  - CPU and storage resources
  - Support

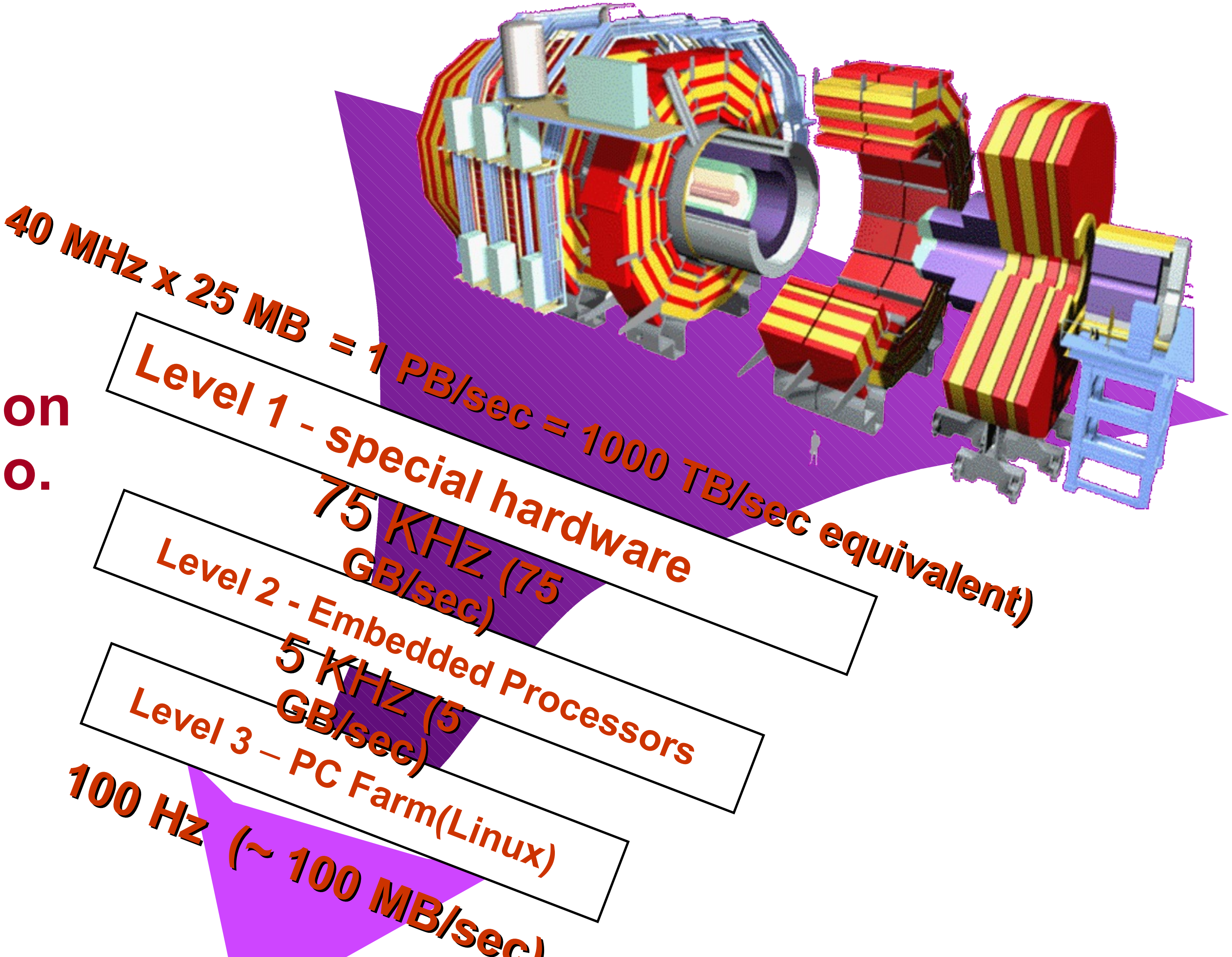
Karlsruher Institut für Technologie

The Worldwide LHC Computing Grid (WLCG)

Andreas Heiss

Data rates of the LHC experiments

data reduction 1/10 Mio.



~ 2 PB per year per experiment (+ simulations)

The Worldwide LHC Computing Grid (WLCG) - the 'fifth experiment'

Karlsruher Institut für Technologie

Memorandum of Understanding

for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid

between

The EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH ("CERN"), an intergovernmental Organization having its seat at Geneva, Switzerland, as the Host Laboratory of the Worldwide LHC Computing Grid, the provider of the Tier0 Centre and the CERN Analysis Facility, and as the coordinator of the LCG project,

on the one hand,

and

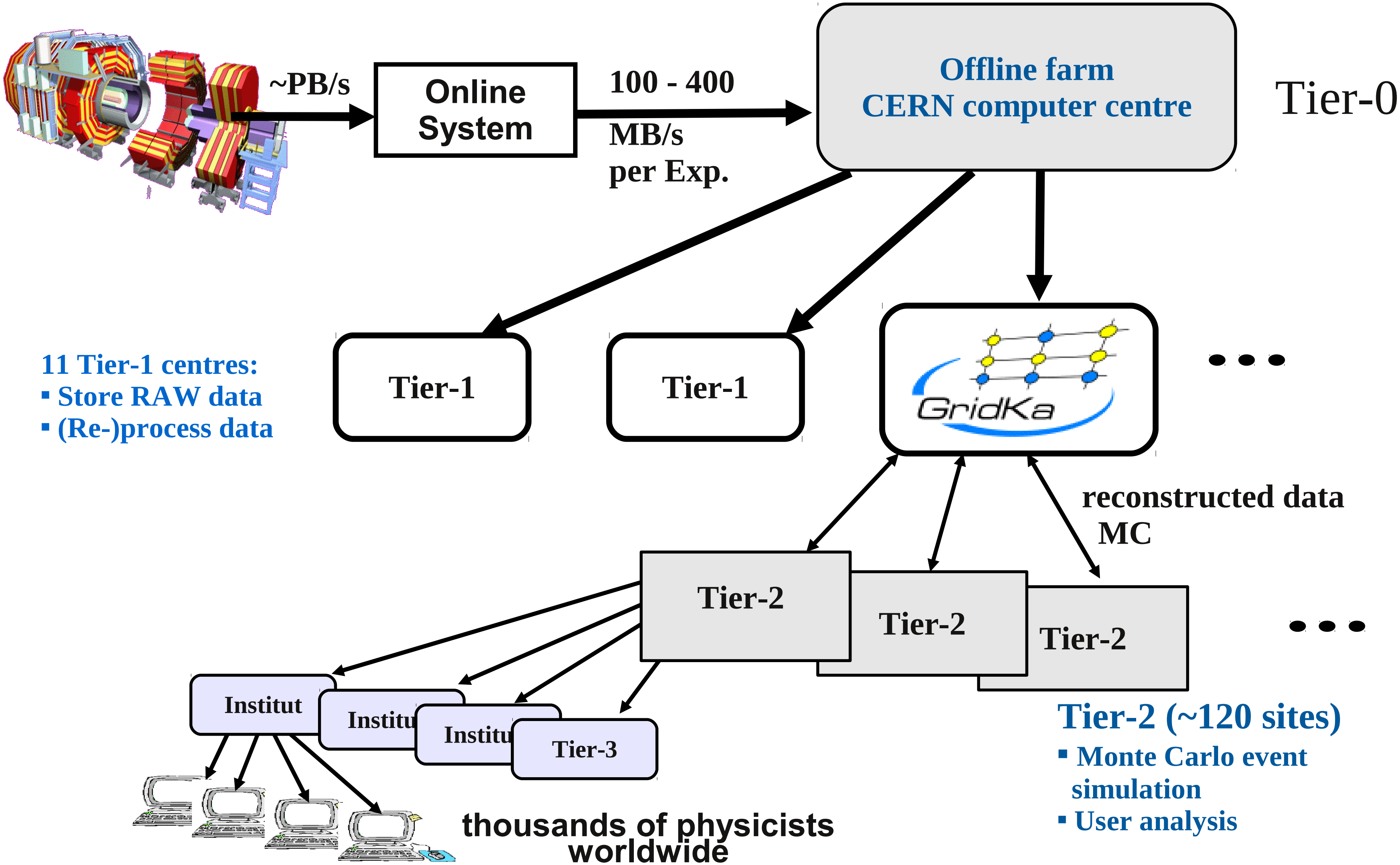
all the Institutions participating in the provision of the Worldwide LHC Computing Grid with a Tier1 and/ or Tier2 Computing Centre (including federations of such Institutions with computer centres that together form a Tier1 or Tier2 Centre), as the case may be, represented by their Funding Agencies for the purposes of signature of this Memorandum of Understanding,

on the other hand,

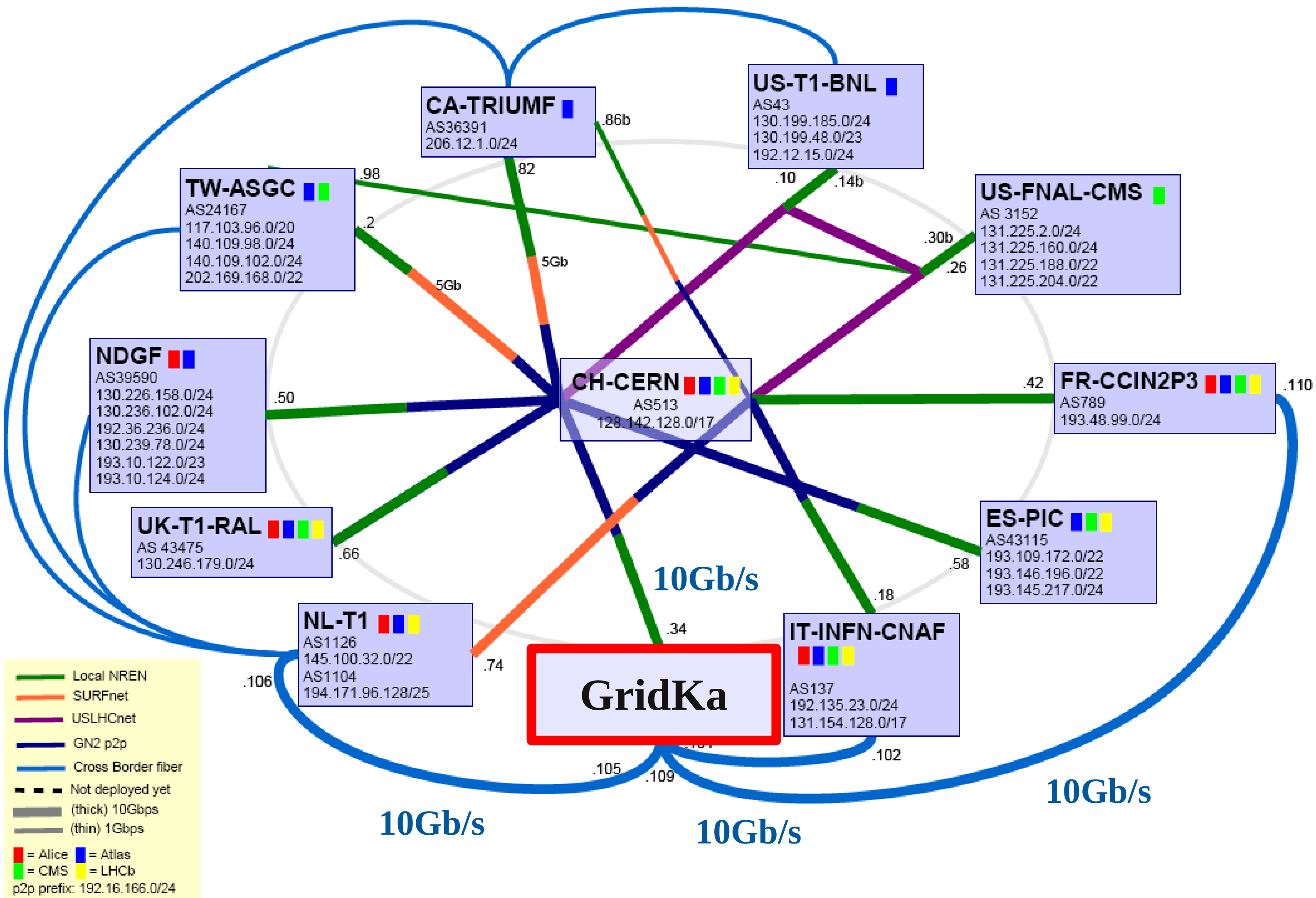
- Participating countries (funding agencies)
- LHC experiments
- Computing and storage resources
- Service levels
- Project organisation and management

The WLCG computing model

Karlsruher Institut für Technologie

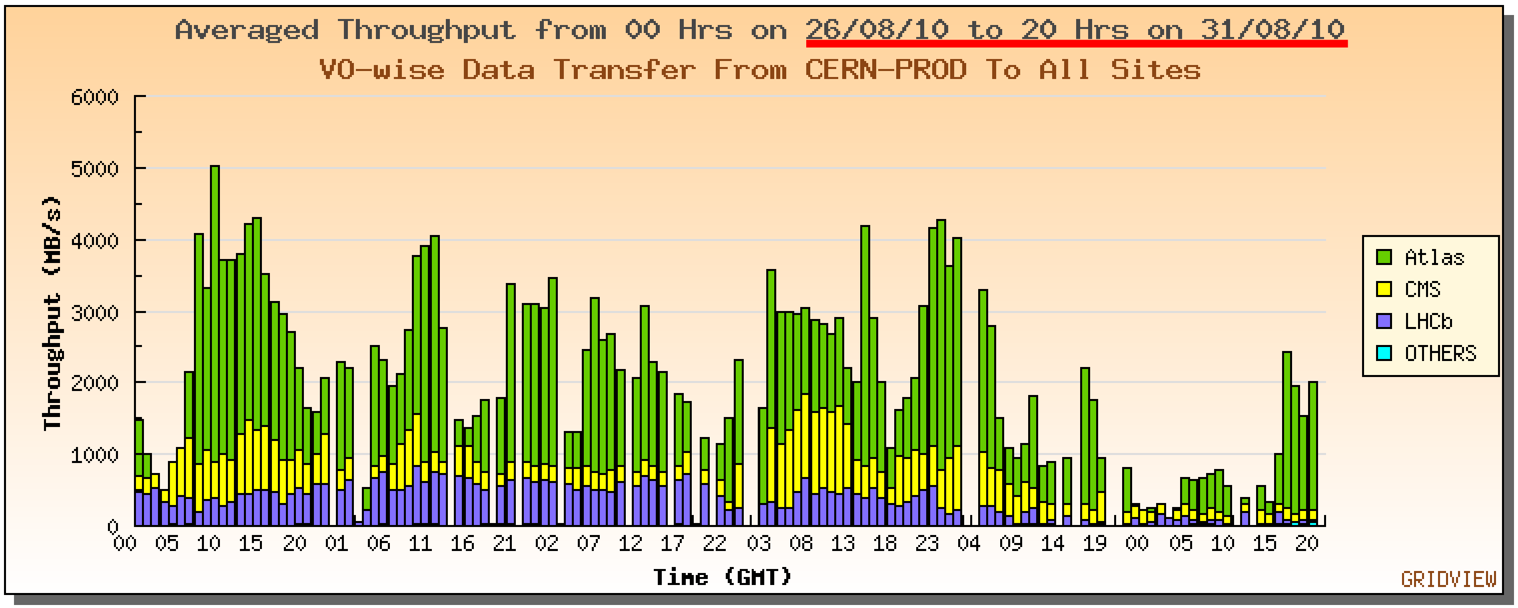


The LHC optical private network



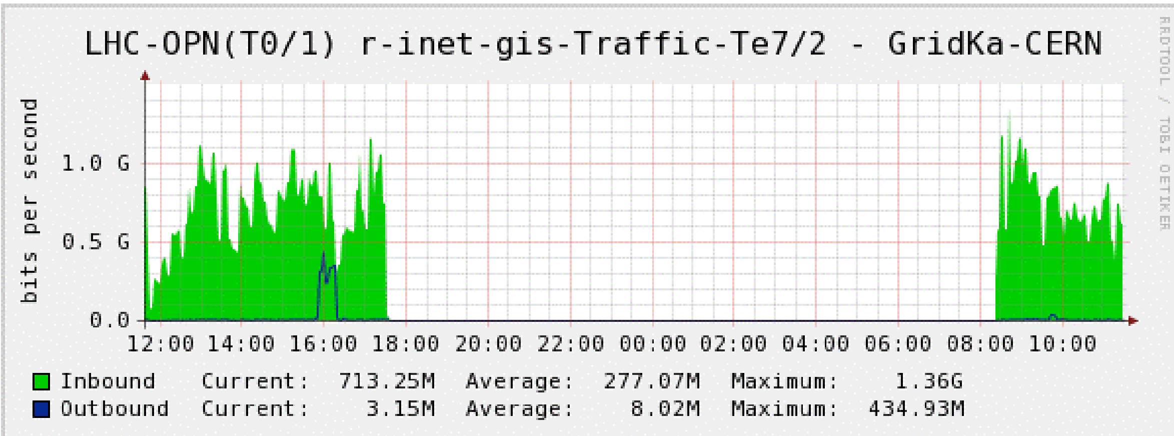
The LHC optical private network

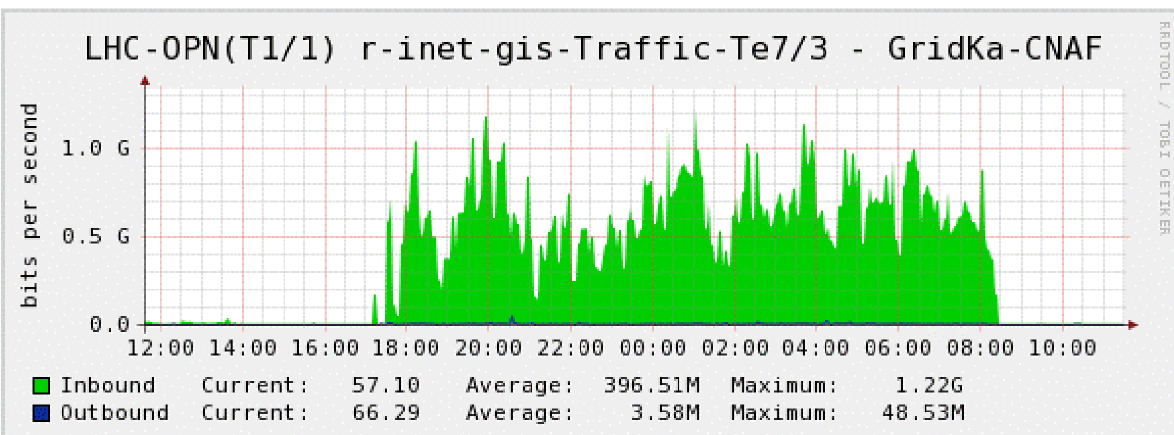
Data rate from CERN to Tier-1 sites



The LHC optical private network

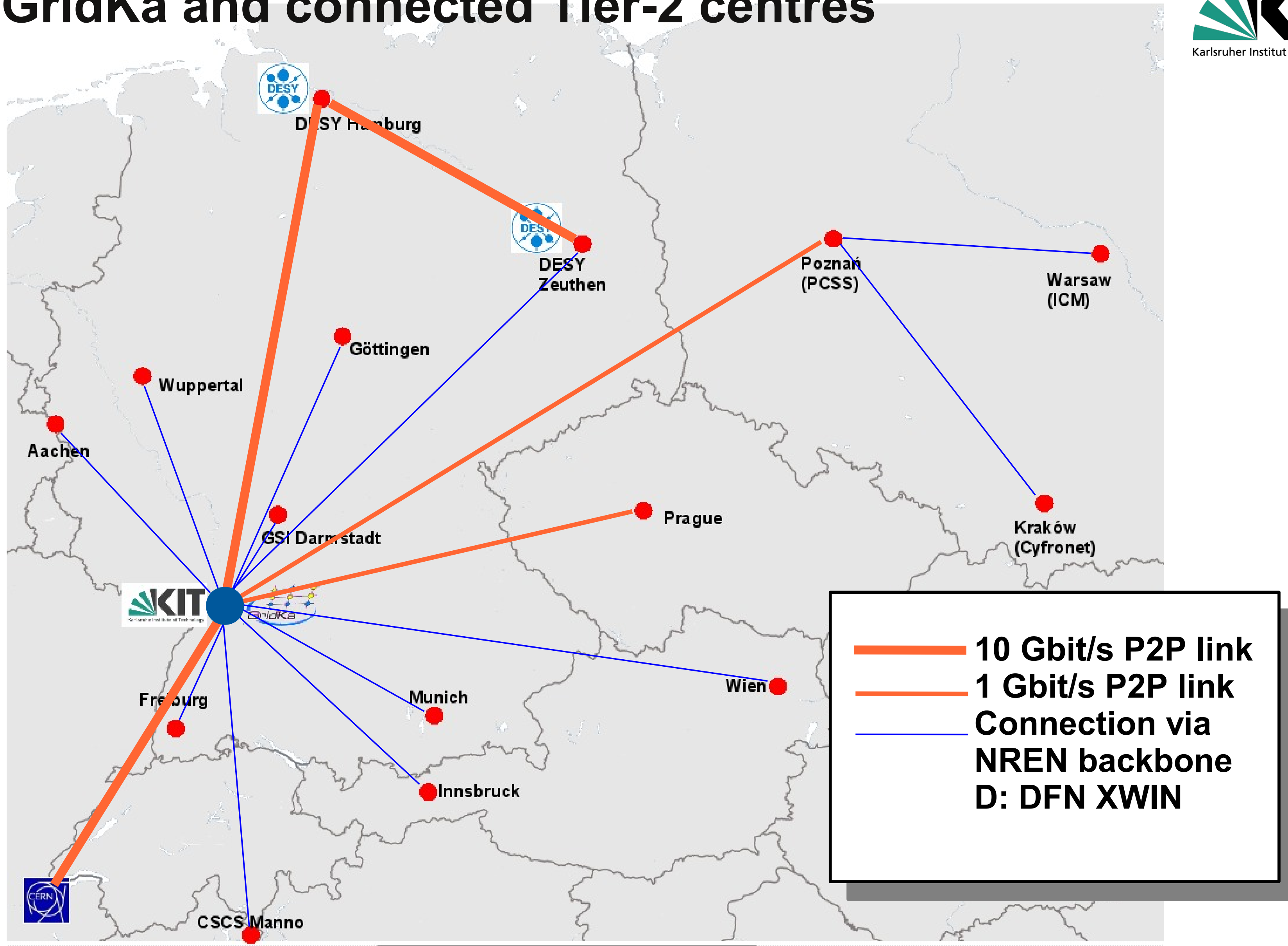
Automatic failover : network failure of the LHCOPN link between CERN and GridKa





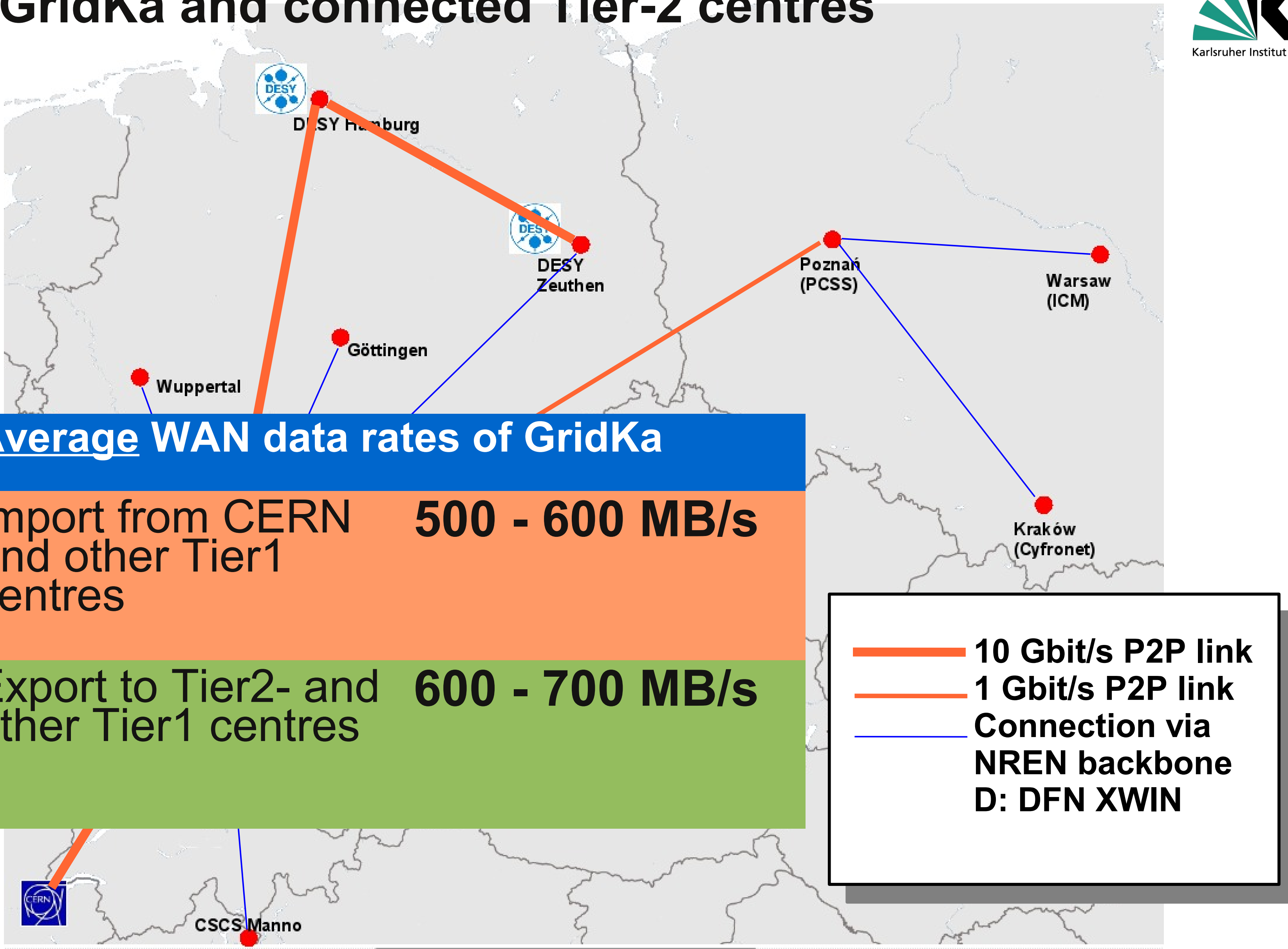
routing of T0-T1 traffic over the backup link via CNAF

GridKa and connected Tier-2 centres



Andreas Heiss

GridKa and connected Tier-2 centres



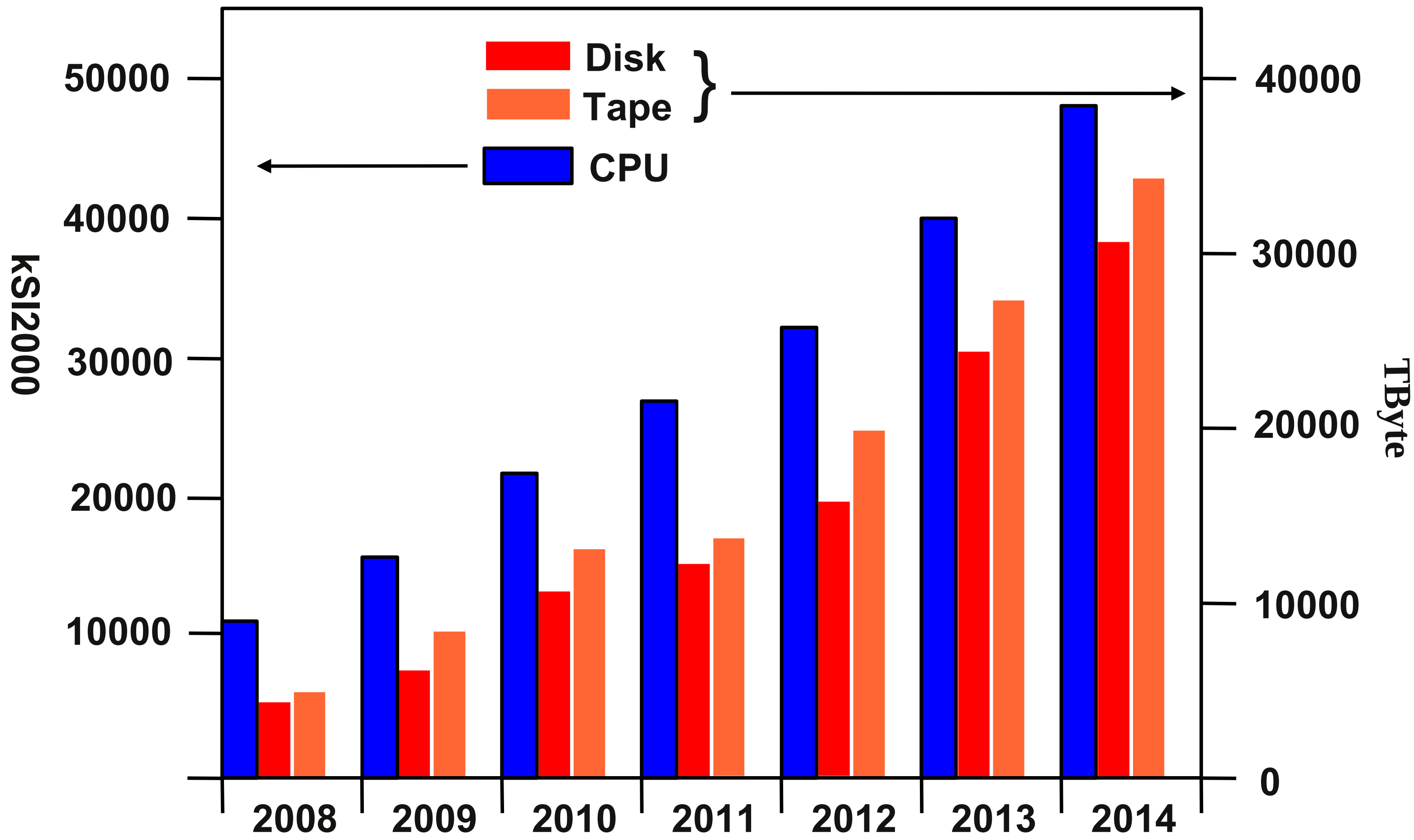
Average WAN data rates of GridKa	
Import from CERN and other Tier1 centres	**500 - 600 MB/s**
Export to Tier2- and other Tier1 centres	**600 - 700 MB/s**

 Andreas Heiss

Resources

Andreas Heiss

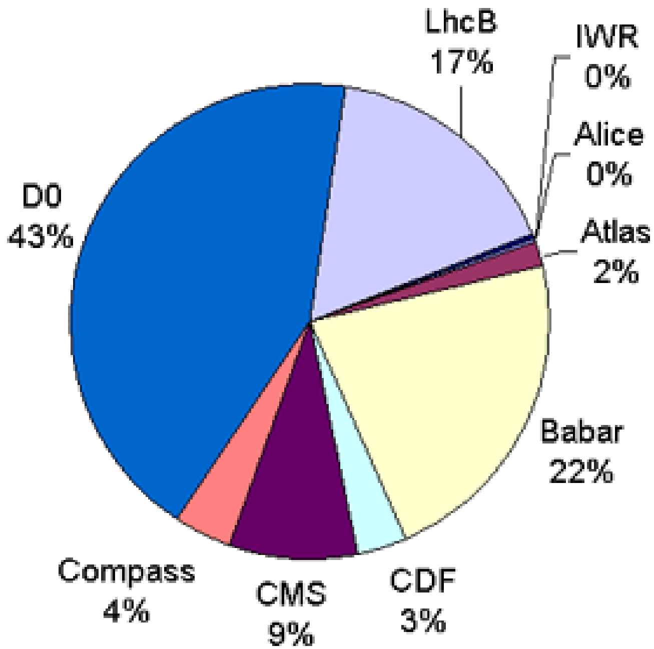
GridKa compute power and storage capacity
(approx. numbers)



2010: ~10000 CPU cores, ~10000 Terabytes disk, >10000 Terabytes tape

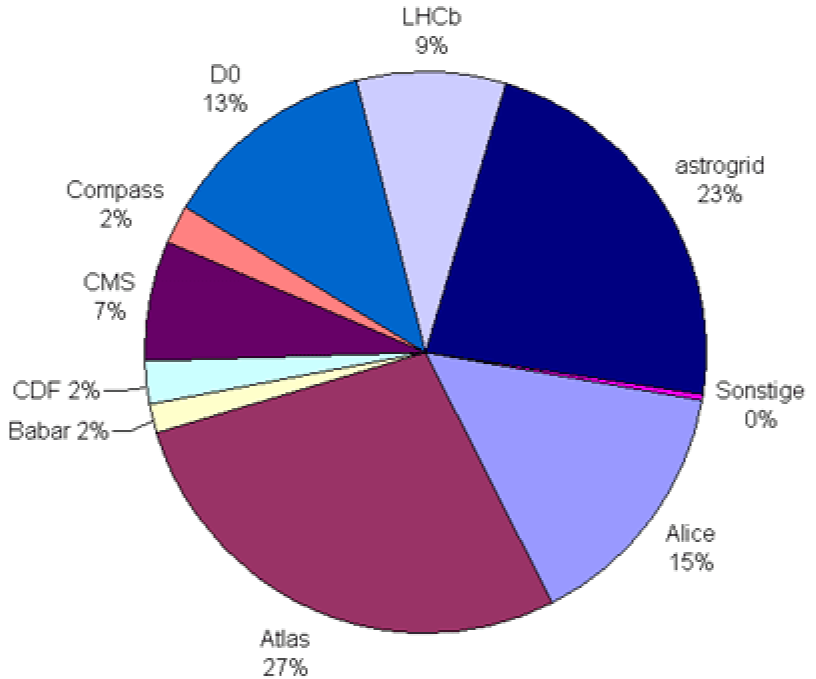
Resource usage: CPU

2003



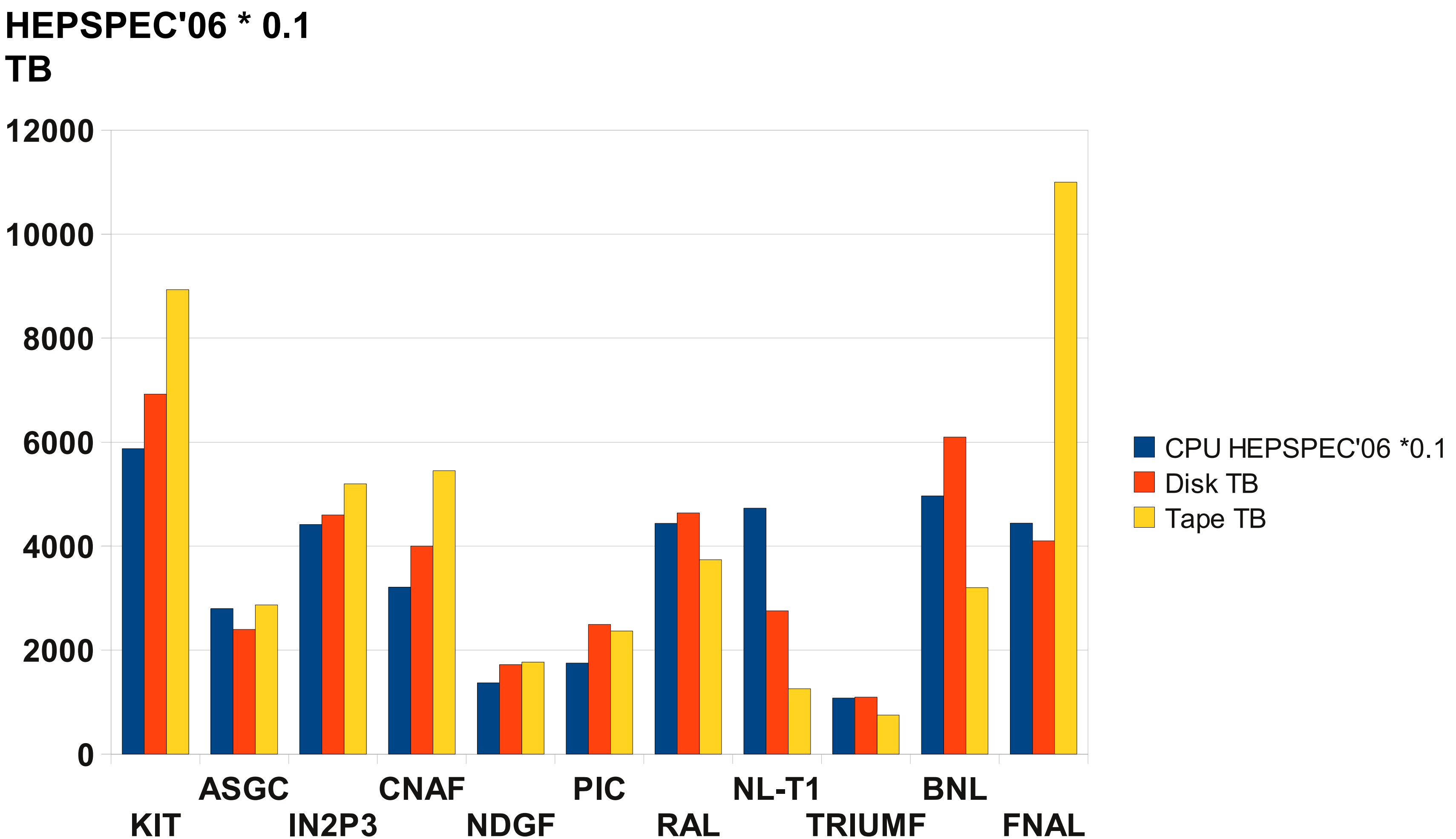
1 980 000 hours
(LHC: 28%)

2009



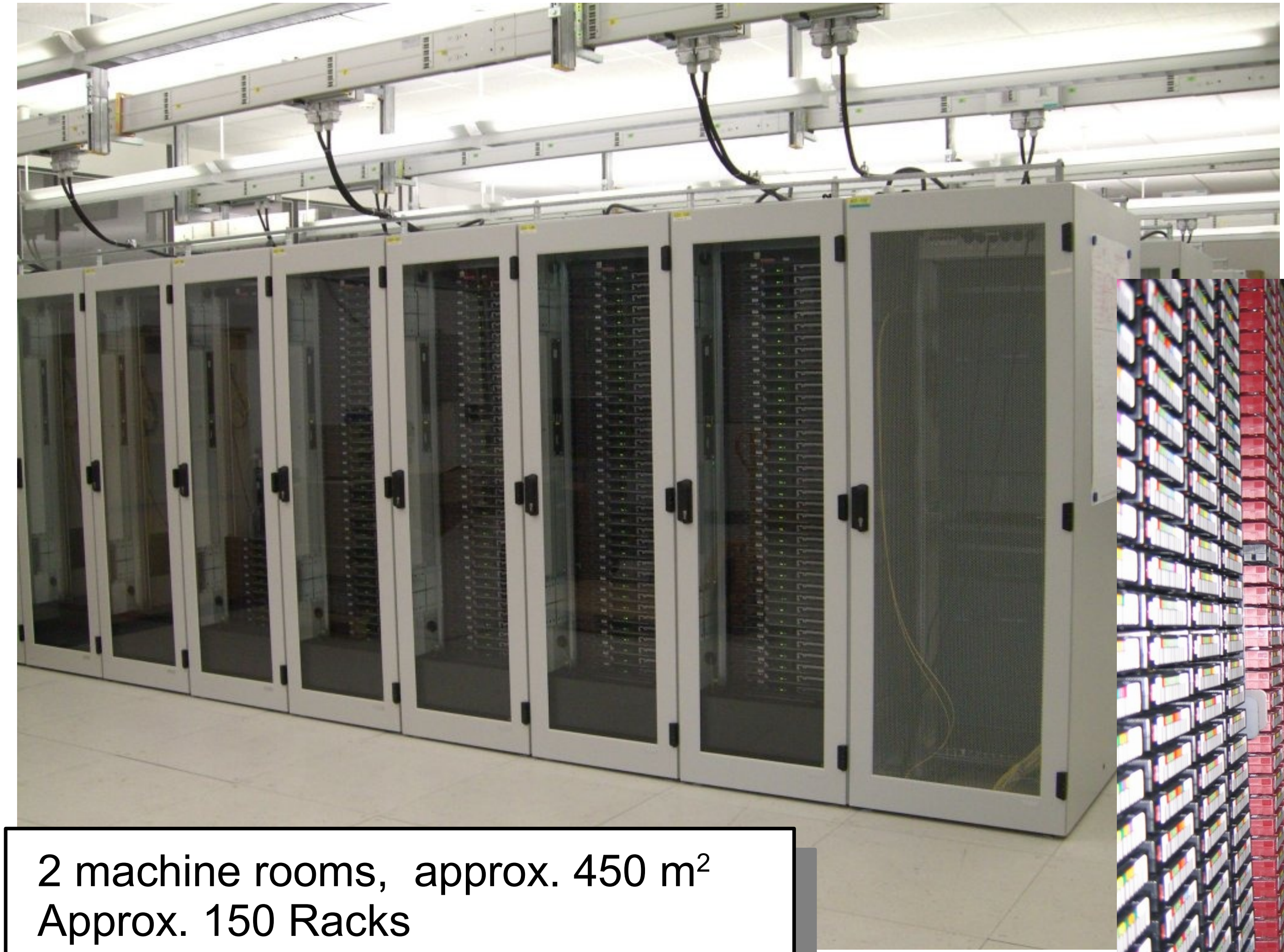
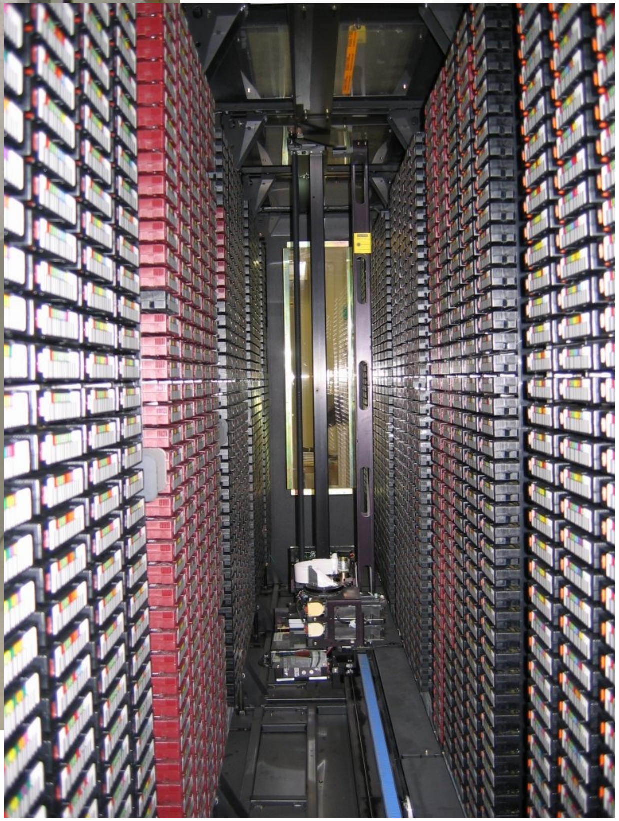
42 770 000 hours
(LHC: 58%)

LHC Tier-1 computing resources

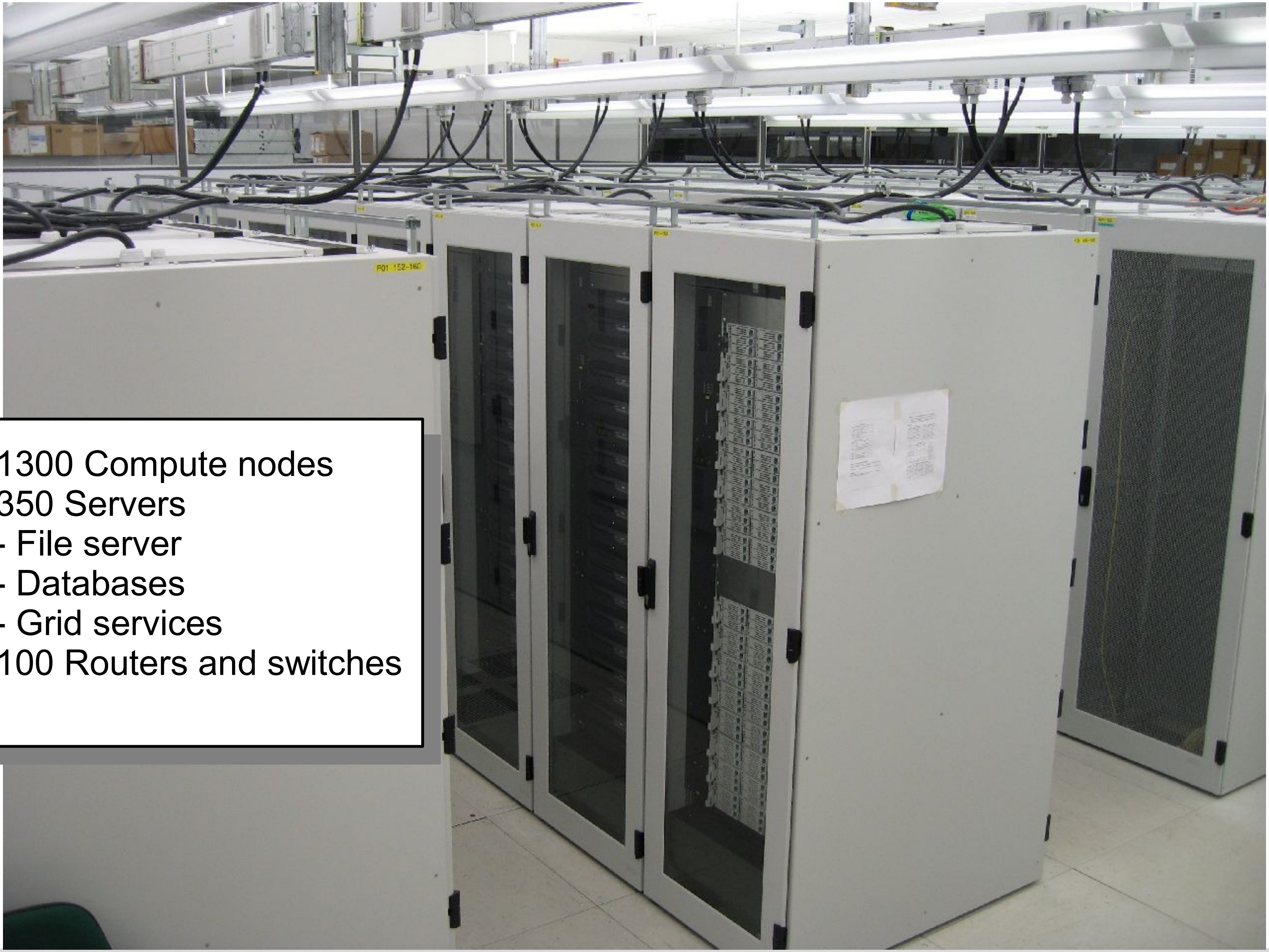


Andreas Heiss

- **Ramp-up once per year in April**
  - *O*(100) compute nodes
  - PetaBytes of disks and many servers
- **Accurate planning is a must!**
  - Infrastructure
    - Power, Cooling
    - Floorspace and racks
    - SAN and LAN
  - Time
    - European procurement procedures: >6 months
    - (Wo)man power

2 machine rooms, approx. 450 m^2
Approx. 150 Racks



> 1300 Compute nodes
~ 350 Servers
- File server
- Databases
- Grid services

> 100 Routers and switches

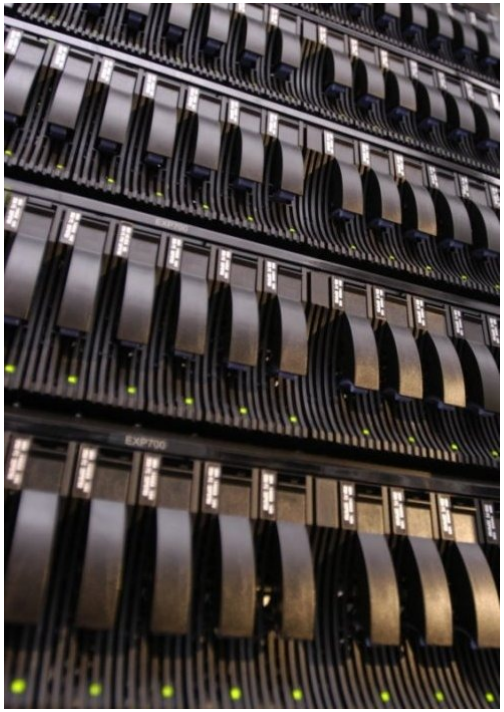
- **Big increase of resources every year**
  - **can result in scaling problems:**
    - Batch system
    - Network / shared file systems
    - Storage systems
    - LAN and SAN
    - Management and monitoring systems

→ Take new hardware into production in several steps

- **Hardware operated for 3-5 years**
  - Mixture of older and newer hardware in production, e.g.
    - 4 core and 8 core compute nodes
    - File servers with 1GE or 10GE LAN
    - File systems of 1TB or 18TB size
    - ...

    } Risk of data access bottlenecks!
  - Replacement of hardware necessary
    - compute nodes → easy
    - disk systems or file servers

    → O(PB) of data has to be copied every year
      - should be transparent to users



Services

Grid services @ GridKa

Local services	
CE	Compute Element: interface to local batch system
sBDII	Information system: publishes information about resources and services
SE, SRM	Storage element, storage resource manager
APEL, DGAS	Accounting
Regional / global services	
RB, WMS	Resource broker, workload management system
BDII	Global Grid information system
LFC	File catalogue: maps between logical file names and physical files in storage elements
FTS	File transfer service: schedules and performes file transfers

Grid services @ GridKa

Local services	
CE	Failure has mostly local impact.
sBDII	
SE, SRM	
APEL, DGAS	
Regional / global services	
RB, WMS	Failure has regional or even global impact. (Tier-2 centres, Tier-1↔ Tier-2 data transfers)
BDII	
LFC	
FTS	

Andreas Heiss

Experiment specific services @ GridKa

Local services	
VOBOX	Runs experiment specific services, e.g. PhEDEx, Alien, ...
Databases	Access to conditions data
Squid	Database cache
Regional / global services	
LFC	Special instance of LFC, data streamed from CERN to GridKa

Services @ GridKa

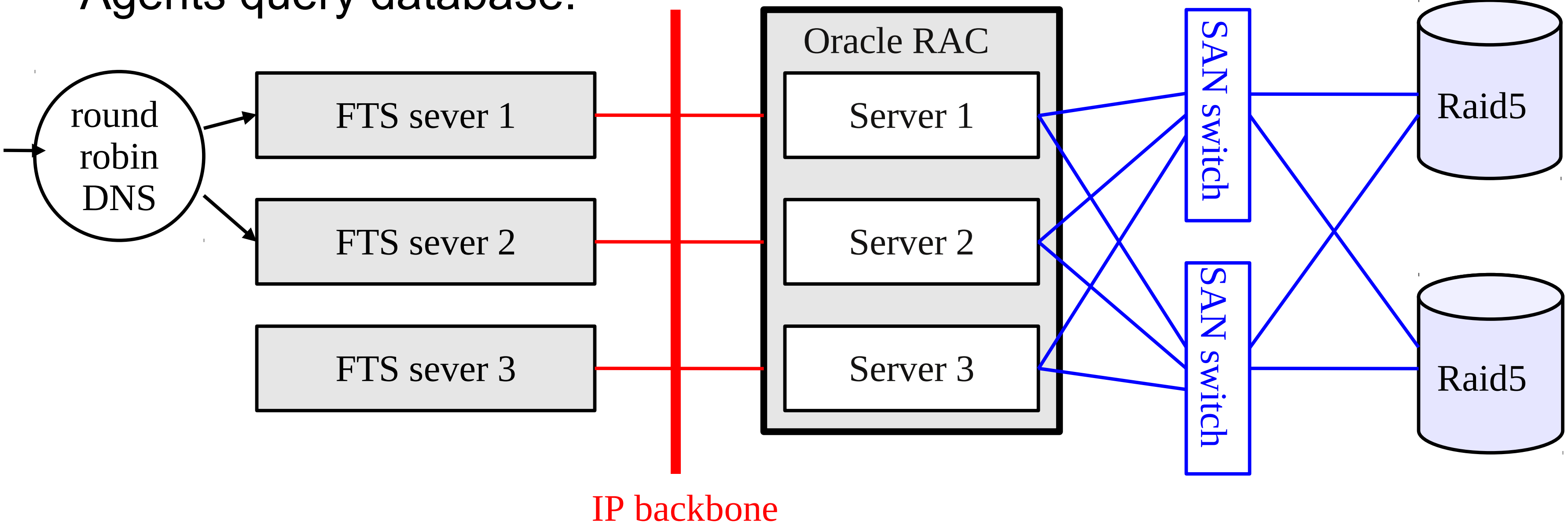
- **Computing models of HEP experiments rely on many different services**

  (e.g. a data transfer involves: FTS, LFC, SRM, SE, VOBOX)
  - **Several single points of failure for each task**
  - **High availability of services is essential**
    - **The total availability is the *product* of the availability of the individual components.**

- **Redundancies of services and service components**
  - **Several instances of services**
    - **e.g. CE, WMS**
    - **Failover mechanism ideally to be implemented in the *client* (if the first does not work, try the second)**
  - **High availability setup of services**
    - **e.g. FTS, LFC**

Example: setup of FTS at GridKa

- FTS servers each run a web service.
- Transfer and VO agents distributed, can be moved to another machine in case of failure.
- Oracle database RAC stores transfer jobs and job status.
- Agents query database.



- FTS servers distributed in two racks
- Oracle RAC distributed in two racks
- All hardware with redundant power supplies

 Andreas Heiss

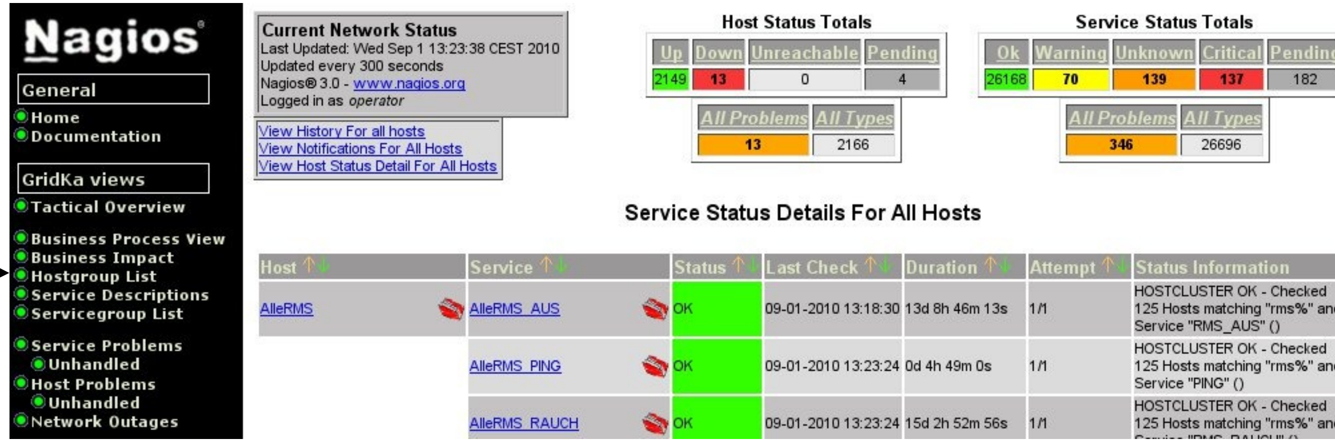
Operations

Operation of a Tier-1 centre

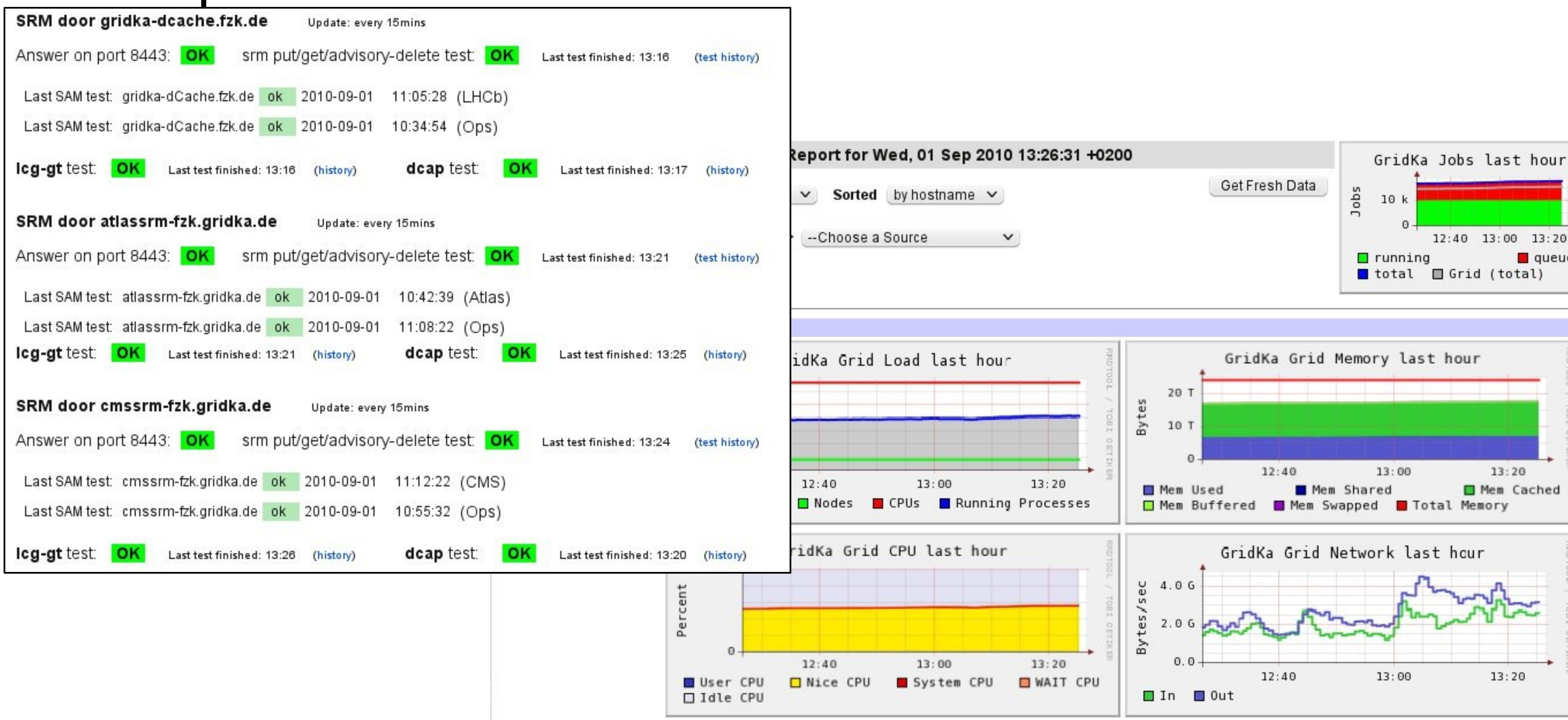
- Management tools
  - OS installation
  - Configuration of OS and services
    - Scalability
    - Administrator mistakes can have large impact

- Monitoring
  - Error and performance monitoring
  - Error condition can be a logic combination of several metrics
  - Definition of alarm conditions for on-call service
    - Avoid false positive alarms

Monitoring



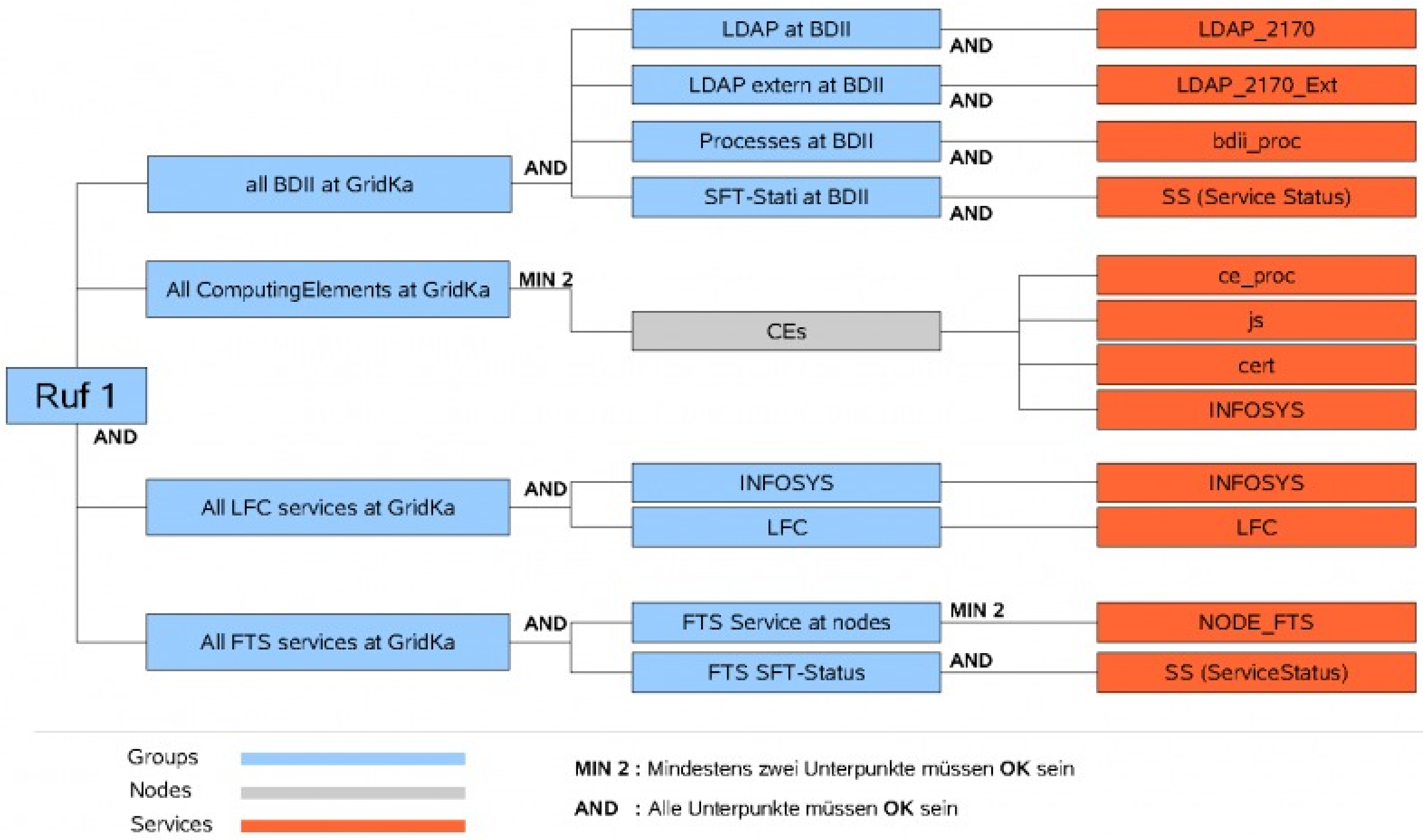
Nagios
General
Home
Documentation
GridKa views
Tactical Overview
Business Process View
Business Impact
Hostgroup List
Service Descriptions
Servicegroup List
Service Problems
Unhandled
Host Problems
Unhandled
Network Outages
Current Network Status
Last Updated: Wed Sep 1 13:23:38 CEST 2010
Updated every 300 seconds
Nagios® 3.0 - www.nagios.org
Logged in as operator
View History For all hosts
View Notifications For All Hosts
View Host Status Detail For All Hosts
Host Status Totals
Up Down Unreachable Pending
2149 13 0 4
All Problems All Types
13 2166
Service Status Totals
Ok Warning Unknown Critical Pending
26168 70 139 137 182
All Problems All Types
346 26696
Service Status Details For All Hosts
Host Service Status Last Check Duration Attempt Status Information
AlleRMS AlleRMS_AUS OK 09-01-2010 13:18:30 13d 8h 46m 13s 1/1 HOSTCLUSTER OK - Checked 125 Hosts matching "rms%" and Service "RMS_AUS" ()
AlleRMS_PING OK 09-01-2010 13:23:24 0d 4h 49m 0s 1/1 HOSTCLUSTER OK - Checked 125 Hosts matching "rms%" and Service "PING" ()
AlleRMS_RAUCH OK 09-01-2010 13:23:24 15d 2h 52m 56s 1/1 HOSTCLUSTER OK - Checked 125 Hosts matching "rms%" and



SRM door gridka-dcache.fzk.de Update: every 15mins
Answer on port 8443: OK srm put/get/advisory-delete test: OK Last test finished: 13:16 (test history)
Last SAM test: gridka-dCache.fzk.de ok 2010-09-01 11:05:28 (LHCb)
Last SAM test: gridka-dCache.fzk.de ok 2010-09-01 10:34:54 (Ops)
lcg-gt test: OK Last test finished: 13:16 (history) dcap test: OK Last test finished: 13:17 (history)
SRM door atlassrm-fzk.gridka.de Update: every 15mins
Answer on port 8443: OK srm put/get/advisory-delete test: OK Last test finished: 13:21 (test history)
Last SAM test: atlassrm-fzk.gridka.de ok 2010-09-01 10:42:39 (Atlas)
Last SAM test: atlassrm-fzk.gridka.de ok 2010-09-01 11:08:22 (Ops)
lcg-gt test: OK Last test finished: 13:21 (history) dcap test: OK Last test finished: 13:25 (history)
SRM door cmssrm-fzk.gridka.de Update: every 15mins
Answer on port 8443: OK srm put/get/advisory-delete test: OK Last test finished: 13:24 (test history)
Last SAM test: cmssrm-fzk.gridka.de ok 2010-09-01 11:12:22 (CMS)
Last SAM test: cmssrm-fzk.gridka.de ok 2010-09-01 10:55:32 (Ops)
lcg-gt test: OK Last test finished: 13:26 (history) dcap test: OK Last test finished: 13:20 (history)
Report for Wed, 01 Sep 2010 13:26:31 +0200
Sorted by hostname
Get Fresh Data
--Choose a Source
GridKa Jobs last hour
Jobs
running queued total Grid (total)
idKa Grid Load last hour
Nodes CPUs Running Processes
GridKa Grid Memory last hour
Bytes
Mem Used Mem Shared Mem Cached Mem Buffered Mem Swapped Total Memory
idKa Grid CPU last hour
Percent
User CPU Nice CPU System CPU WAIT CPU Idle CPU
GridKa Grid Network last hour
Bytes/sec
In Out

Monitoring

Example: on-call alarm condition for Grid services



Andreas Heiss

Security

- GridKa is responsible for valuable data which must be secure.
- Security issues could result in unauthorised access to and abuse of resources, e.g. a large number of compute nodes with a 'pretty good' internet connection.

→ IT security is of high importance!

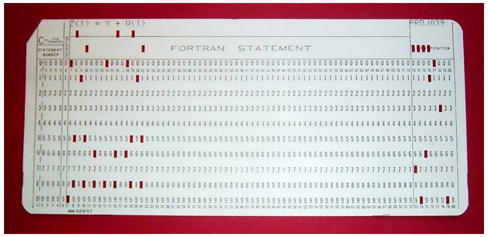
- GridKa security measures include:
  - Intrusion detection on all 'exposed' systems,
    e.g. CEs, login nodes etc.
  - Methods to immediately block users on all systems
  - Experts in computer security and forensics on site
  - Collaboration with other Grid sites: an incident there could be a threat to GridKa also! (e.g. stolen ssh-key or Grid certificate)
  - ...

(Future) challenges

Andreas Heiss

(Future) challenges

- Long term (>20 years) storage of experiments' data
  - Copy data to new storage media types
  - Things to consider:
    - Lifetime of media?
    - Special software required?
      (protocols, file systems, ...)

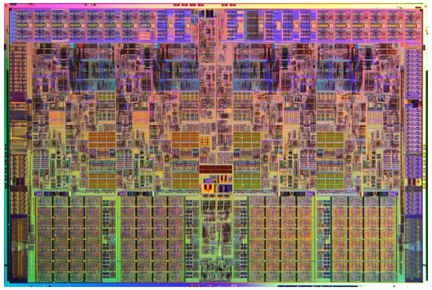


- Still evolving computing models
  - IT techniques improve every year
  - Experiments change computing models based on their experiences and (new) technical possibilities.
    - but need to access and compute old data as well

Andreas Heiss

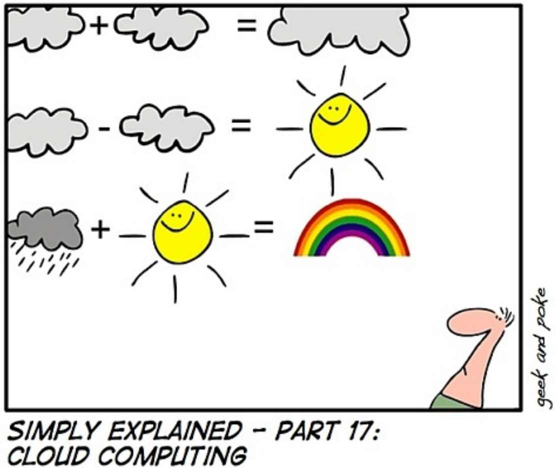
(Future) challenges

- Technology trends, e.g.
  - Trend to more CPU cores instead of higher clock frequencies
    - Higher LAN bandwidth required
    - More simultaneous file accesses
    - Multithreaded jobs?



  - Network bandwidth grows faster than local I/O
    - 100Gbit/s WAN on the horizon
    - New possibilities arise → influence on computing models?

- New computing paradigms arise
  - e.g. Cloud computing



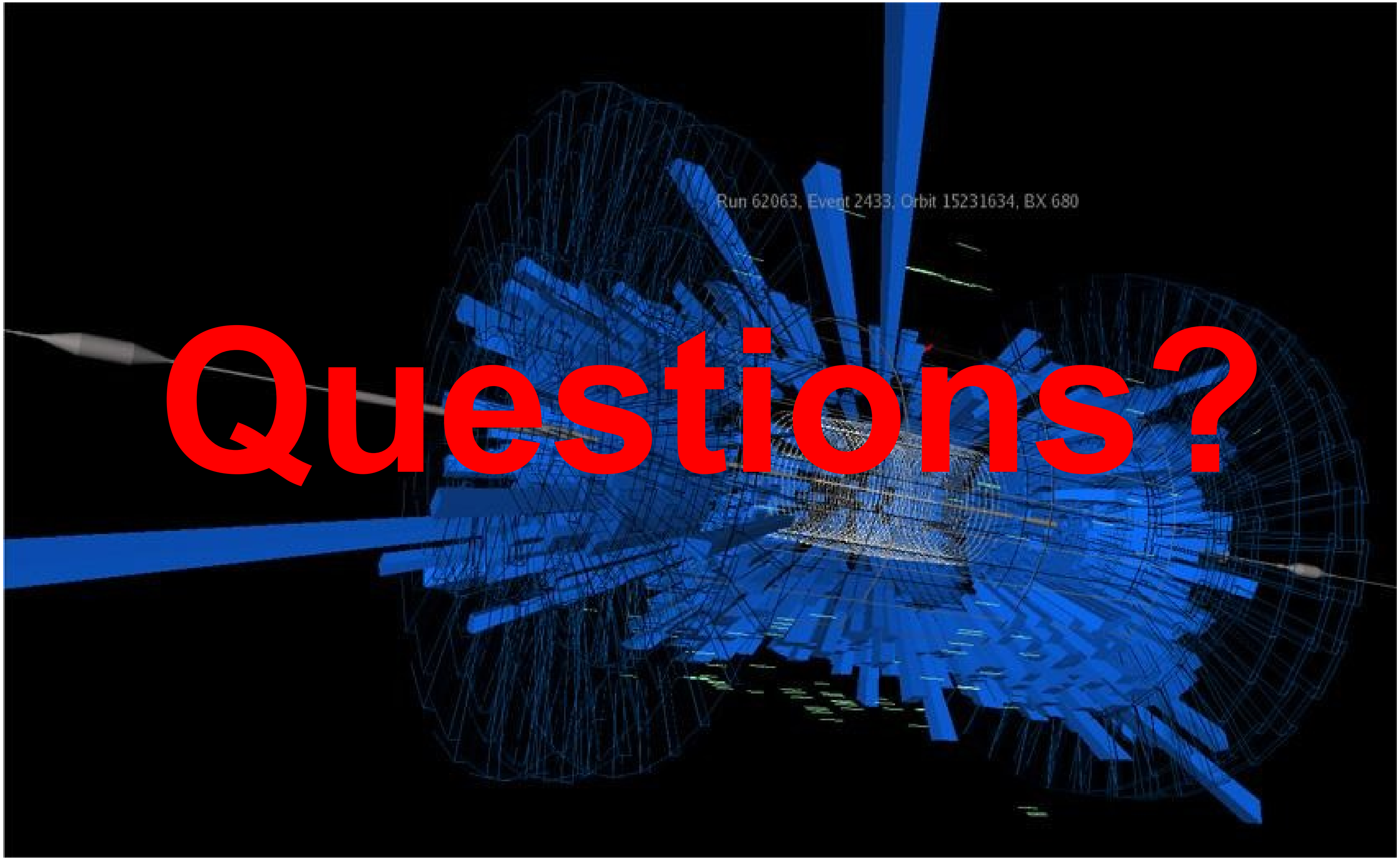
Andreas Heiss

Outlook

- GridKa project phase 3 has started.
  - The LHC is taking data!
  - Are we finished? *NO!*

- There's still a lot to do:
  - New technologies to be tested and deployed.
  - New services to be installed.
  - Computing models will change.
  - Keep GridKa state-of-the-art!

Andreas Heiss


Questions?
Run 62063, Event 2433, Orbit 15231634, BX 680