The Tier-1 centre GridKa

Dr. Andreas Heiss
KIT north campus
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 participates in national and international projects and working groups:
  - Test setups
  - R&D
  - CPU and storage resources
  - Support
The Worldwide LHC Computing Grid (WLCG)
Data rates of the LHC experiments

- Level 1 - special hardware: 40 MHz \times 25 \text{MB} = 1 \text{PB/sec} = 1000 \text{TB/sec equivalent}

- Level 2 - Embedded Processors: 75 \text{KHZ (75 GB/sec)}

- Level 3 - PC Farm (Linux): 5 \text{KHZ (5 GB/sec)}

- Data reduction: 1/10 Mio.

- Total: \approx 2 \text{PB per year per experiment (+ simulations)}
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,
The WLCG computing model

~PB/s

Online System

100 - 400 MB/s per Exp.

Offline farm
CERN computer centre

Tier-0

11 Tier-1 centres:
- Store RAW data
- (Re-)process data

Tier-1

Tier-1

Tier-2

Tier-2

Tier-2 (~120 sites)
- Monte Carlo event simulation
- User analysis

reconstructed data
MC

thousands of physicists worldwide

Institut

Institut

Institut

Tier-3

...
The LHC optical private network
The LHC optical private network
Data rate from CERN to Tier-1 sites

Averaged Throughput from 00 Hrs on 26/08/10 to 20 Hrs on 31/08/10
VO-wise Data Transfer From CERN-PROD To All 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
Andreas Heiss

13 Gbit/s P2P link
1 Gbit/s P2P link
Connection via NREN backbone
D: DFN XWIN

GridKa and connected Tier-2 centres
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

- 10 Gbit/s P2P link
- 1 Gbit/s P2P link
- Connection via NREN backbone
- D: DFN XWIN

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

- D0: 43%
- Lhcb: 17%
- CMS: 9%
- CDF: 3%
- Babar: 22%
- Compass: 4%
- IWR: 0%
- Alice: 0%
- Atlas: 2%

1 980 000 hours (LHC: 28%)

2009

- LHCb: 9%
- D0: 13%
- Compass: 2%
- CMS: 7%
- CDF: 2%
- Babar: 2%
- Atlas: 27%
- Astrogrid: 23%
- Alice: 15%
- Sonstige: 0%

42 770 000 hours (LHC: 58%)
LHC Tier-1 computing resources

HEPSPEC'06 * 0.1
TB

0 2000 4000 6000 8000 10000 12000

- CPU HEPSPEC'06 *0.1
- Disk TB
- Tape TB

KIT ASGC CNAF NDGF PIC RAL NL-T1 TRIUMF BNL FNAL
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 m²
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
  - ...
  
- Replacement of hardware necessary
  - compute nodes → easy
  - disk systems or file servers
    → $O(\text{PB})$ of data has to be copied every year
    - should be transparent to users

Risk of data access bottlenecks!
Services
## Grid services @ GridKa

### Local services

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>Compute Element: interface to local batch system</td>
</tr>
<tr>
<td>sBDII</td>
<td>Information system: publishes information about resources and services</td>
</tr>
<tr>
<td>SE, SRM</td>
<td>Storage element, storage resource manager</td>
</tr>
<tr>
<td>APEL, DGAS</td>
<td>Accounting</td>
</tr>
</tbody>
</table>

### Regional / global services

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RB, WMS</td>
<td>Resource broker, workload management system</td>
</tr>
<tr>
<td>BDI</td>
<td>Global Grid information system</td>
</tr>
<tr>
<td>LFC</td>
<td>File catalogue: maps between logical file names and physical files in storage elements</td>
</tr>
<tr>
<td>FTS</td>
<td>File transfer service: schedules and performs file transfers</td>
</tr>
<tr>
<td>Grid services @ GridKa</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td></td>
</tr>
</tbody>
</table>

**Local services**

- CE: Compute Element
- sBDII: Information system
- 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
- FTS: File transfer service

---

Failure has mostly local impact.

Failure has regional or even global impact.

(Tier-2 centres, Tier-1 ↔ Tier-2 data transfers)
## Experiment specific services @ GridKa

<table>
<thead>
<tr>
<th>Local services</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VOBOX</td>
<td>Runs experiment specific services, e.g. PhEDEx, Alien, ...</td>
</tr>
<tr>
<td>Databases</td>
<td>Access to conditions data</td>
</tr>
<tr>
<td>Squid</td>
<td>Database cache</td>
</tr>
<tr>
<td>Regional / global</td>
<td></td>
</tr>
<tr>
<td>services</td>
<td></td>
</tr>
<tr>
<td>LFC</td>
<td>Special instance of LFC, data streamed from CERN to GridKa</td>
</tr>
</tbody>
</table>
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
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

Example: on-call alarm condition for Grid services
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
(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
(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
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!
Questions?