GridKa – The German Tier-1 Computing Centre: Status, Strategy & Future Plans

Forschungszentrum Karlsruhe GmbH
Hauptabteilung Informations- und Kommunikationstechnik
Hermann-von-Helmholtz-Platz 1
D-76344 Eggenstein-Leopoldshafen

Dr. Holger Marten

http://grid.fzk.de
Content

1. HEP Computing @ GridKa
2. Cluster Installation
3. A rack-based cooling system
4. Data management
5. Grid & R&D Activities
1. HEP Computing @ GridKa
History I


08/2000: Launching Workshop for EU-DataGrid in Marseille (France)
- F. Gagliardi: “Germany is under-represented; additional activities are welcome”

10/2000: German physicists formulate requirements for a LHC Tier-1 Centre
- Forschungszentrum Karlsruhe presents design ideas and cost estimates
The Large Hadron Collider Project - 4 detectors

ATLAS

CMS

LHCb

Alice

DESY Seminar, May 12, 2003
The LHC multi-Tier Computing Model

Tier 0
- Tier 0 Centre at CERN
- working groups
- virtual organizations

Tier 1
- CERN Tier 1
- USA (Fermi, BNL)
- UK (RAL)
- France (IN2P3)
- Germany (FZK)
- Italy (CNAF)

Tier 2
- Tier 2 (Uni-CCs, Lab-CCs)
- Lab x
- Lab y
- Lab z

Tier 3
- Tier 3 (Institute computer)
- Uni e
- Uni f
- Uni g
- Uni h
- Uni i

Tier 4
- Tier 4 (Desktop)
- Lab a
- Lab b
- Lab c

Germany
- Uni a
- Uni b
- Uni c
- Uni d
- Uni e

……

Tier 1

DESY Seminar, May 12, 2003
History II

05/2001: BaBar-D contract signed; FZK and Univ. Bochum buy a common Linux-Cluster

07/2001: German HEP communities send “Requirements for a Regional Data and Computing Centre in Germany (RDCCG)”

11/2001: FZK presents planning and cost estimates for RDCCG

12/2001: Launching committee establishes RDCCG project structure

04/2002: first prototype of RDCCG goes into operation

07/2002: RDCCG renamed to “Grid Computing Centre Karlsruhe, GridKa”
What is the Grid Computing Centre Karlsruhe?

German Regional Centre for Grid Computing with focus
- LHC Tier-1 at 2008+
- test environment for LHC (**ALICE, ATLAS, CMS, LHCb**) now!
- production environment for non-LHC (**BaBar, CDF, D0, Compass**) now!
- test environment for Grid R&D (**CrossGrid, DataGrid, LCG, ...**) now!
- user support, services, education & training **a.s.a.p.**
- environment for other sciences (**bio-informatics, astrophysics,...**) later

**It’s a project with 41 user groups from 19 German institutions**
GridKa Project Structure

- **Project Leader & Deputy**
  H. Marten, M. Kunze

- **Overview Board**
  controls execution & financing, arbitrates in case of conflicts
  includes 6 representatives of HEP

- **Technical Advisory Board**
  defines technical requirements
  includes 18 representatives of HEP
GridKa Overview Board (OB)

**Chairman**
R. Maschuw (FZK board of directors)

**Representative of BMBF**
J. Richter

**Project Leader & Deputy**
H. Marten, M. Kunze (FZK)

**Head of FZK Computing Department**
K.-P. Mickel (FZK)

**Chairman of TAB**
P. Malzacher (GSI)

**Representatives of KET & KHK**
R.-D. Heuer, (DESY), P. Braun-Munzinger (GSI)

2 Representatives of LHC/non-LHC each
L. Köppke (U Mainz) for Atlas
P. Mättig (U Wuppertal) for CDF/D0
T. Müller (TU Karlsruhe) for CMS
M. Schmelling (MPI Heidelberg) for Alice & LHCb
B. Spaan (TU Dresden) for BaBar & Compass
GridKa Technical Advisory Board (TAB)

**Chairmen**
K.-P. Mickel (FZK), P. Malzacher (GSI)

**Project Leader & Deputy**
H. Marten, M. Kunze (FZK)

**Representatives of KET & KHK**
R.-D. Heuer, (DESY) P. Braun-Munzinger (GSI)

**Representative of DESY**
R. Mankel (DESY)

**8 representatives of LHC experiments**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Representative 1</th>
<th>Representative 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>K. Schwarz (GSI Darmstadt)</td>
<td>P. Malzacher (GSI Darmstadt)</td>
</tr>
<tr>
<td>Atlas</td>
<td>G. Duckeck (LMU München)</td>
<td>L. Köppke (U Mainz)</td>
</tr>
<tr>
<td>CMS</td>
<td>G. Quast (TU Karlsruhe)</td>
<td>F. Raupach (RWTH Aachen)</td>
</tr>
<tr>
<td>LHCb</td>
<td>M. Schmelling (MPI Heidelberg)</td>
<td>J. Blouw (U Heidelberg)</td>
</tr>
</tbody>
</table>

**8 representatives of non-LHC experiments**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Representative 1</th>
<th>Representative 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>BaBar</td>
<td>H. Lacker (TU Dresden)</td>
<td>M. Steinke (RU Bochum)</td>
</tr>
<tr>
<td>CDF</td>
<td>T. Müller (TU Karlsruhe)</td>
<td>K. Rinnert (TU Karlsruhe)</td>
</tr>
<tr>
<td>Compass</td>
<td>F.-H. Heinsius (U Freiburg)</td>
<td>L. Schmitt (TU München)</td>
</tr>
<tr>
<td>D0</td>
<td>D. Wicke (U Wuppertal)</td>
<td>Ch. Zeitnitz (U Mainz)</td>
</tr>
</tbody>
</table>
German users of GridKa

- 19 institutions
- 41 user groups
- ~ 350 scientists

Aachen (4) ▲
Bielefeld (2) ●
Bochum (2) ●
Bonn (3) ●
Darmstadt (1)▲
Dortmund (1) ●
Dresden (2) ●
Erlangen (1) ●
Frankfurt (1) ●
Freiburg (2) ●
Heidelberg (1)▲ (6)●
Karlsruhe (2) ●
Mainz (3) ●
Mannheim (1) ●
München (1)▲ (5)●
Münster (1) ●
Rostock (1) ●
Siegen (1) ●
Wuppertal (2) ●

University
▲ other research institutions
() Number of working groups

August 2002
GridKa – Planned yearly hardware upgrades

- CPU (kS/i95)
- Disk (TByte)
- Tape (TByte)
Organisational Structure

HIK (K.-P. Mickel)

Zentralabtlg.

Sekretariat

DASI
Datendienste
Hochleistungsrechnen
Infrastruktur

(R. Kupsch)

HLR
Grid-Computing Infrastruktur und Service

(F. Schmitz)

GES
Grid-Computing und e-Science

(H. Marte)

NiNa
Netzinfrastruktur und Netzanwendungen

(K. -P. Mickel)

NiNa
PC/BK
PC-Betreuung und Bürokommunikation

(A. Lorenz)

Repro
Reprografie

H. Marte

GridKa

Zentralabteilung und Sekretariat: IT-Innovationen, Accounting, Billing, Budgetierung, Ausbildungskoordination, sonstige zentrale Aufgaben

DESY Seminar, May 12, 2003
2. Cluster Installation & Design
Support for multiple experiments I

• experiments ask for RedHat 6.2, 7.2, 7.1 or Fermi Linux, SuSE 7.2 ... in different environments
• experiments ask for Grid (Globus, EDG,...), batch & interactive login

• split the CPUs into 8 parts ?
  - would be administrative challenge
  - likely, whole environment would be busy only part time
• reconfigure for each experiment ?
  - non-LHC experiments produce and analyse data all the time
  - who should define a time schedule ?
  - what about other sciences ?
Experiment Specific Software Server

8x dual PIII for Alice, Atlas, BaBar, ..., each with
• 2 GB ECC RAM, 4x 80 GB IDE-Raid5, 2x Gbit Ethernet
• Linux & basic software on demand:
  RH 6.2, 7.1.1, 7.2, Fermi-Linux, SuSE 7.2

Used as
• Development environment per experiment
• Interactive login & Globus gatekeeper per experiment

• Basic admin (root) by FZK
• Specific software installation by experiment admin
Dedicated Grid LAN Backbone

Extreme Black Diamond 6808
- redundant power supply
- redundant management board
- 128 Gbit/s back plane
- max. 96 Gbit ports

Upgrade to Black Diamond 6816 today
- 256 Gbit/s back plane
- max. 192 Gbit ports
Compute Nodes

188x dual PIII, PIV, each with
- 1, 1.26, 2.2 GHz
- 1 GB ECC RAM
- 40 GB HDD IDE
- 100 Mbit Ethernet

Total numbers:
- 7.5 TB local disk
- 188 GByte RAM
- $R_{\text{peak}} > 500$ GFlops

- RedHat 7.2
- OpenPBS
- automatic installation with NPACI Rocks
Cluster Installation, Monitoring & Management

- **scalability**: many nodes to install and maintain (ca. 2000)
- **heterogeneity**: different (Intel-based?) hardware over time
- **consistency**: software must be consistent on all nodes
- **manpower**: administration by few persons only

This is for Administrators, not for a Grid Resource Broker
Philosophies for Cluster Management

- **scalability:**
  - hierarchical instead of pure central management
  - combined push and pull for management information
  - info & event handling via separate management network

- **heterogeneity:** rpm instead of disk cloning

- **consistency:** distribute software from a central service

- **manpower:** automatise as much as you can
Architecture for Scalable Cluster Administration

Cabinet 1

Nodes C₁

Manager C₁

Cabinet 2

Nodes C₂

Manager C₂

Cabinet n

Nodes Cₙ

Manager Cₙ

Private Compute Network

Management Network

Master

Public Net

Naming scheme: C02-001...064; F01-003,...
Installation - NPACI Rocks with FZK extensions

- DHCP-server for Managers C1...Cn
- RH kickstart incl. IP conf. for Managers
- rpm to Managers

reinstall all nodes in < 1 h

http://rocks.npaci.edu
System Monitoring with Ganglia

- also installed on fileservers
  - CPU usage
  - Bytes I/O
  - Packets I/O
  - disk space
  - ...
- and published on the Web

http://ganglia.sourceforge.net
System Monitoring with Ganglia - Combined Push-Pull

- Ganglia daemon on each node
- info via multicast
- no routing

**subnet**

- Manager C₁
- Nodes C₁
- Nodes C₂
- Nodes Cₙ

**Private Compute Network**

**Management Network**

**Ganglia Master**

write into round robin DB
300 kB / node

**publish to Web**
Cluster Management with Nagios - Combined Push-Pull

- Analyse data
- Handle local events

Manager C₁

Nodes C₁

Nodes C₂

Nodes Cₙ

Private Compute Network

Management Network

Nagios Master

GPL

http://www.nagios.org

Ping

SNMP

syslog

report

• Analyse data
• Handle events

Publish to Web
Software Installation & Central Services

Software Installation Service (NPACI Rocks)
http://rocks.npaci.edu

Open PBS Server

System Monitoring (Ganglia Cluster Toolkit)
http://ganglia.sourceforge.net

Globus MDS, LDAP,... (Globus 2.0)

System Management
http://www.nagios.org

User Statistics (home made)
3. The cooling problem
Infrastructure

We all build Linux clusters...

... and what about cooling?

4 kW/m²

12 kW/m²
Closed rack-based cooling system -
A common development of FZK and Knürr

- 19’’ technique
- 36 units height usable
- 70x 120 cm floor space
- 10 kW cooling
- redundant DC fans
- temperature controlled
- CPU shutdown
- internal smoke detector
Closed rack-based cooling system -
Estimated cost reduction > 70% compared to air conditioning
... and that’s the reality
4. GridKa data management
Online Storage

- 59 TB brutto
- 45 TB net capacity
- ~ 500 disk drives
- mixed IDE, SCSI, FC
- ext3 & ufs file systems

**DAS:** 2.6 TB brutto, 7.2k SCSI 120 GB, attached to SUN Enterprise 220 R

**SAN:** 2x 5.1 TB brutto, 10k FC 73,4 GB, IBM Fast500

**NAS:** 42.2 TB brutto, 19x IDE-System, dual PIII 1.0/ 1.26 GHz,
    dual 3Ware Raid Controller, 16x 5.4k IDE 100/ 120/ 160 GB

**SAN-IDE:** 3.8 TB brutto, 2 Systems, 12x 5.4k IDE 160 GB, driven by Linux-PC
Disk Storage & Management – does it scale?

• >600 automount operations per second for 150 processors
• measured IDE NAS throughput
  - >150 MB/s local read (2x RAID5 + RAID0)
  - 30-40 MB/s w/r .... over NFS
  ... but < 10 MB/s with multiple I/O and multiple users
• 300 jobs write to a single NAS box
• Linux file system limit 2 TB
• disk volumes of >50 TB with flexible volume management desirable
• mature system needed now !

We will test gfs & gpfs for Linux
Scheme of (Disk) Storage

Cluster Nodes (clients)

Tests with Linux Server & FC/IDE successful

SAN (shared w. FZK)

SAN goes commodity!

Cluster Nodes (clients)

SAN (shared w. FZK)

Disk Subsystems

Tape Drives

multiple Gbit

Grid backbone

Fileserver IDE NAS 1.

GridKa

DESY Seminar, May 12, 2003
GridKa Tape System

IBM 3584
- ~ 2400 slots LTO Ultrium
- 100 GB/tape
- 106 TB native available
- 8 drives, 15 MByte/s each
- Backup/Archive with Tivoli Storage Manager

FZK SAN
Discussion of Tape Storage Management

- HPSS, TSM, SAM-FS, gen. HSM ... do exist
  - for vendor specific file systems
  - on vendor specific disk systems
  - for vendor specific tape systems

- File systems, data management & mass storage are strongly coupled
- Almost every HEP experiment has own data management philosophy
Tape Storage & Data Management under discussion

- GridKa disk storage
- robotics
- tape error handling
- tape recycling

DataGrid
Castor
backup
SAM / JIM
dCache
TSM
5. Grid & R&D Activities
Grid installation & testing

- dedicated CrossGrid / DataGrid installations available at GridKa
- EDG 1.3.8, 1.4.x installed
- ALICE: own Grid solution Alien, job-submit via Globus
- ATLAS: job-submit via Globus
- CDF/ D0: own Grid solution “JIM”
- GridKa is prototype for D0- Regional Analysis Centre
- BaBar plans data analysis via EDG; dedicated gatekeeper installed
- LCG-0 runs since April 24, 2003
GermanGrid CA

- delivers x509 certificates for German Grid research activities of
  - Alice, Atlas, CMS, LHCb
  - BaBar, CDF, Compass, D0
  - CrossGrid, DataGrid, LCG
    - already delivered certificates to GridLab

- Certificate Policy and Certification Practice Statement available

http://grid.fzk.de/ca

Certificates accepted by DataGrid- and CrossGrid;
Certificates delivered by globus.org not accepted!
GridKa – Focus on R&D

Marcel Kunze  Division Leader GES
Ariel Garcia  EU CrossGrid & EU DataGrid testbeds
Marcus Hardt  EU CrossGrid & EU DataGrid testbeds
Holger Marten  EU CrossGrid, WP2 manager “Grid Programming Environment”
Gerardo Ganis  LCG: Integration of middleware into the Root system
Kathrin Paschen  LCG: Grid technology
Peer Hasselmeyer  LCG, Grid Middleware Testbeds & deployment
Andreas Heiss  Optimization of Dataflows in clusters, Infiniband
Ulrich Schwickerath  Optimization of Dataflows in clusters, Infiniband
GridKa – Focus on planning, testing, installation, running

Holger Marten  Division Leader GIS, Project Leader GridKa
Manfred Alef   Login Server, Security, File Server, Linux, Infrastructure
Peer Hasselmeyer  LCG, Grid Middleware Testbeds & deployment
Bruno Hoeft  LAN/WAN Network
Axel Jaeger  Infrastructure, Cluster Management
Melanie Knoch  Globus, batch, user accounts
Andre Kube  batch, Infrastructure
Ingrid Schäffner  Globus, Web, user accounts, certificates
Bernhard Verstege  Cluster Management, Linux
Jos van Wezel  Data Management, File Server, Linux

Ursula Epting (NiNa)  GermanGrid CA
+ a few persons from other divisions
We appreciate the continuous interest and support by the Federal Ministry of Education and Research, BMBF.