The GridKa Installation for HEP Computing

Forschungszentrum Karlsruhe GmbH
Central Information and Communication Technologies Department
Hermann-von-Helmholtz-Platz 1
D-76344 Eggenstein-Leopoldshafen

Holger Marten

http://grid.fzk.de
Helmholtz Foundation of German Research Centres (HGF)

- 15 German research centres
- 24,000 employees
- largest German scientific organization

- 6 main research areas
  - Structure of Matter
  - Earth & Environment
  - Traffic & Outer Space
  - Health
  - Energy
  - Key Technologies
Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft
Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Forschungszentrum Karlsruhe

• 40 institutes and divisions
• 3,500 employees
• 13 research programs for
  - Structure of Matter
  - Earth & Environment
  - Health
  - Energy
  - Key Technologies
• many close collaborations with TU Karlsruhe
HIK provides institutes of the Research Centre with state-of-the-art high performance computers and IT solutions for each purpose.

vector computers, parallel computers, Linux Clusters, workstations, ~2500 PCs,
online storage, tape robots, networking infrastructure, printers and printing services, central software, user support,....

About 90 persons in 7 departments
Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Organisational Structure

HIK
(K.-P. Mickel)

Zentralabtlg.

Sekretariat

DASI
Datendienste
Anwendungen
Systemüberwachung
Infrastruktur
(R. Kupsch)

HLR
Hochleistungsrechnen
(F. Schmitz)

GIS
Grid-Computing
Infrastruktur
und
Service
GridKa
(H. Marten)

GES
Grid-Computing
und
e-Science
Competence
Centre
GridKa
(M. Kunze)

NiNa
Netzinfrastruktur
und
Netzanwendungen
(K. -P. Mickel)

PC/BK
PC-Betreuung
und
Bürokomunikation
(A. Lorenz)

Repro
Reprografie
(G. Dech)

Zentralabteilung und Sekretariat: IT-Innovationen, Accounting, Billing, Budgetierung, Ausbildungskoordination, sonstige zentrale Aufgaben
Grid Computing Centre Karlsruhe - The mission

German Tier-1 Regional Centre for 4 LHC HEP-Experiments
- 2001-04 test phase
- 2005-07 main set-up phase
- 2007+ production phase

German Computing Centre for 4 non-LHC HEP-Experiments
- 2001-04+ production environment for BaBar, CDF, D0, Compass
Regional Data and Computing Centre Germany, Requirements

For LHC (Alice, Atlas, CMS, LHCb) + BaBar, Compass, D0, CDF

2001-2004: Test Phase  
2005-2007: LHC Setup Phase  
2007+: Operation  

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (kSI95)</td>
<td>1</td>
<td>10</td>
<td>25</td>
<td>60</td>
<td>150</td>
<td>325</td>
<td>900</td>
</tr>
<tr>
<td>disk (TByte)</td>
<td>7</td>
<td>45</td>
<td>113</td>
<td>210</td>
<td>440</td>
<td>850</td>
<td>1500</td>
</tr>
<tr>
<td>tape (Tbyte)</td>
<td>7</td>
<td>111</td>
<td>211</td>
<td>350</td>
<td>800</td>
<td>2000</td>
<td>3700</td>
</tr>
</tbody>
</table>

▲ starts in 2001!  
+ services .... + other sciences
Organization of GridKa

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

• **Overview Board**
  controls execution & financing, arbitrates in case of conflicts

• **Technical Advisory Board**
  defines technical requirements

It’s a project with 41 user groups from 19 German institutions
German users of GridKa

- 19 institutions
- 41 user groups
- ~350 scientists

Universities:
- 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)

Other research institutions:
- Forschungszentrum Karlsruhe
- in der Helmholtz-Gemeinschaft

---

August 2002
The GridKa Installation
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 machine 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 ?
Support for multiple experiments II

Strategy as starting point:

• Compute Nodes = general purpose, shared resources
  - GridKa Technical Advisory Board agreed to RedHat 7.2

• Experiment-specific software server
  - “arbitrary” development environments at the beginning
  - pure software servers and/or Globus gatekeepers later-on
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
Grid LAN Backbone

Extreme Black Diamond 6808
- redundant power supply
- redundant management board
- 128 Gbit/s back plane
- max. 96 Gbit ports
- currently 80 ports available
Compute Nodes

124x dual PIII, each with
- 1 GHz or 1.26 GHz
- 1 GB ECC RAM
- 40 GB HDD IDE
- 100 Mbit Ethernet

Total numbers:
- 5 TB local disk
- 124 GByte RAM
- $R_{peak} > 270$ 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
Public Net

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

subnet

Nodes $C_1$

Manager $C_1$

subnet

Nodes $C_2$

Manager $C_2$

subnet

Nodes $C_n$

Manager $C_n$

Management Network

Private Compute Network

- DHCP-server for Managers $C_1...C_n$
- RH kickstart incl. IP conf. for Managers
- rpm to Managers

Master

Public Net

http://rocks.npaci.edu

reinstall all nodes in < 1 h
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

Manager C₁

Manager C₂

Manager Cₙ

Ganglia Master

write into round robin DB
300 kB / node

publish to Web

Request

Report

Private Compute Network

Management Network
Cluster Management with Nagios - Combined Push-Pull

- analyse data
- handle local events

Manager C₁

Nodes C₁

subnet

ping
SNMP

syslog
report

Management Network

Private Compute Network

Manager C₂

Nodes C₂

Manager Cₙ

Nodes Cₙ

Nagios Master

• analyse data
• handle events

GPL

http://www.nagios.org

publish to Web
Combining Nagios with Tivoli Enterprise Console?

GridKa Cluster

- Nagios Master

TEC

- SMS
- Mail
- inventory DB
- Remedy workflow

Tivoli

- vector computer
- parallel computer
- file server
- Linux Cluster
- STK robots
Infrastructure

We all want to build a Linux cluster...

... do we have the cooling capacity?
Closed rack-based cooling system -
A common development of FZK and Knürr

- 19” technique
- 38 units height usable
- 70x 120 cm floor space
- 10 kW cooling
- redundant DC fans
- temperature controlled
CPU shut-down

Diagram:
- CPU
- disk
- internal heat exchanger
- front
- rear
Closed rack-based cooling system -
Estimated cost reduction > 70% compared to air conditioning
Build a cluster of clusters (or a Campus Grid?)
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
Available Disk Space for HEP: 36 TByte net
Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Scheme of (Disk) Storage

Cluster Nodes (clients)

SAN

Disk Subsystems

Tests with Linux Server & FC/IDE successful

SAN goes commodity!

Grid backbone

Fileserver IDE NAS 1.4

2nd Large Scale Cluster Computing Workshop, FermiLab, October 21-22, 2002
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
• 150 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
Available Tape Space for HEP: 106 TByte native
GridKa Tape System

FZK Tape 1

FZK Tape 2

FZK SAN

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
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
Tape Storage & Data Management under discussion
Summary I - The GridKa installation

- Gbit backbone
- 250 CPUs  + 130 PIV this week  + ~60 PIV until April 2003
- 8 experiment specific server
- 35 TB net disk  + 40 TB net until April 2003
- 100 TB tape  + 110 TB until April 2003 (or on demand)
- a few central servers  Globus, batch, installation, management, ...
- WAN Gbit-test  Karlsruhe-Cern
- FZK-Grid-CA
- crossgrid  --  Data testbed

... exclusively for HEP and Grid Computing
Summary II

There is a whole bunch of questions....

Grid, security, gfs, gpfs, Castor, dCache, stability, reliability, HSM,
certification, hundreds of users, commodity, low cost storage,
OGSA, authorisation, networks, high throughput, ....

.... and a pragmatic solution:

• start with existing technologies
• analyse - learn - evolve
The Federal Ministry of Education and Research, BMBF, considers the construction of GridKa and the German contribution to a World Wide Grid as a national task.