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)

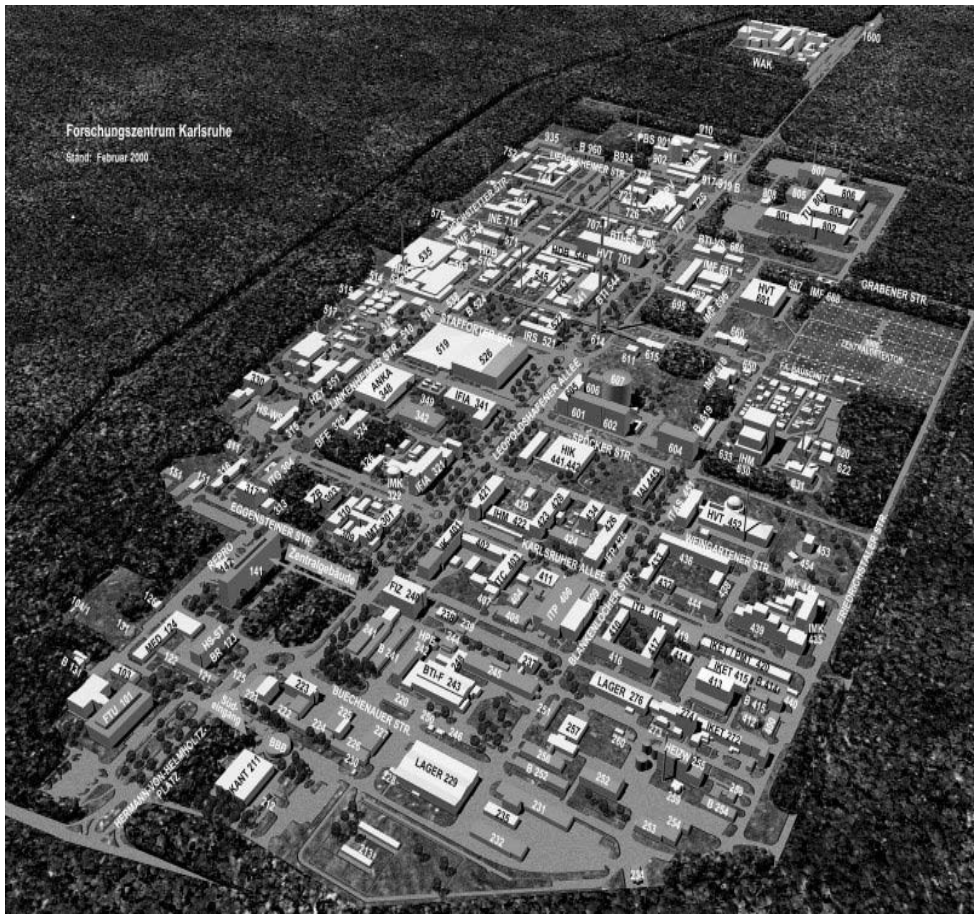
- **15 German research centres**
- **24.000 employees**
- **largest German scientific organization**

- **6 main research areas**
  - **Structure of Matter**
  - **Earth & Environment**
  - **Traffic & Outer Space**
  - **Health**
  - **Energy**
  - **Key Technologies**

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

GridKa

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**

Central Information and Communication Technologies Dept.
(Hauptabteilung Informations- und Kommunikationstechnik, HIK)

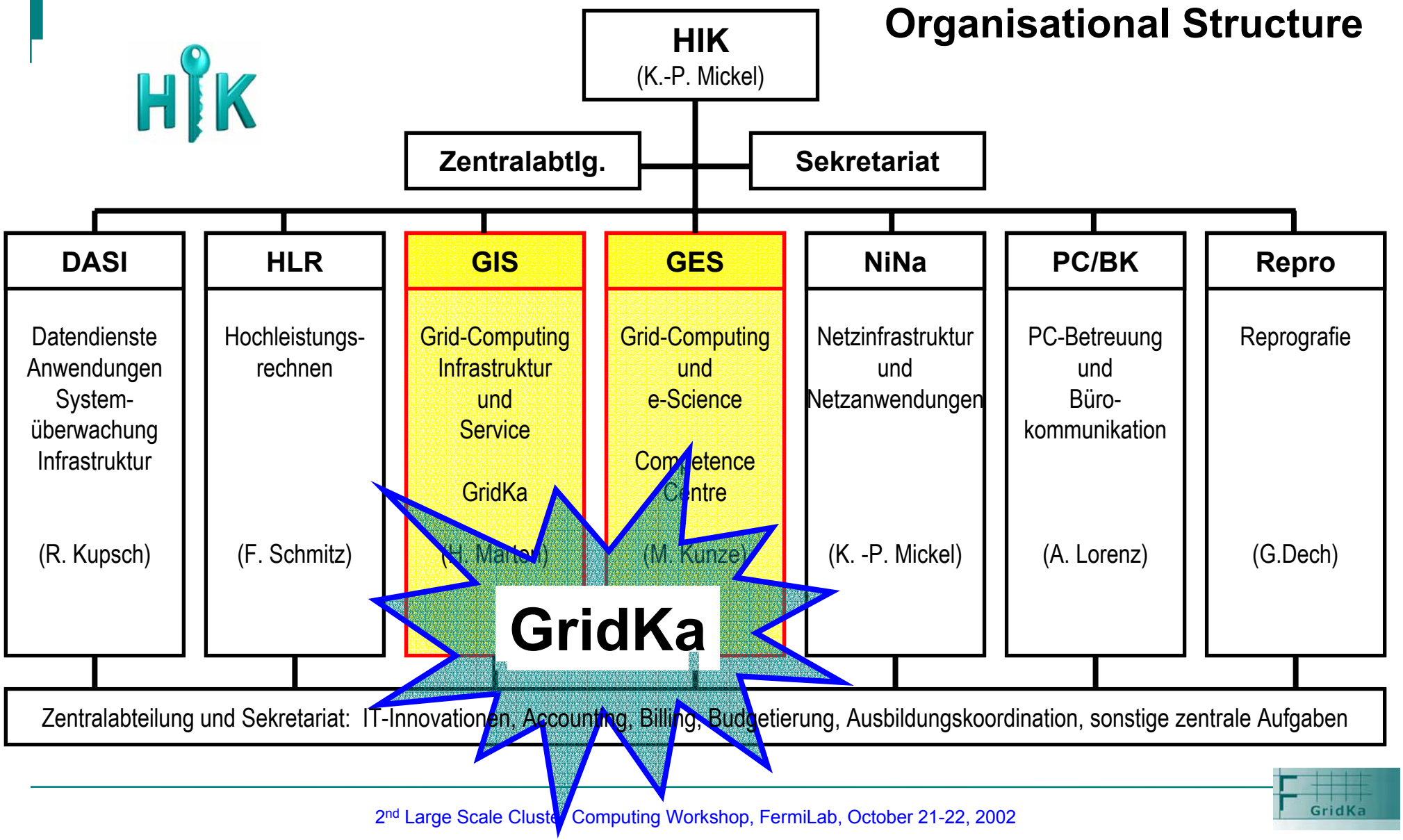
Your Key to Success

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

Organisational Structure



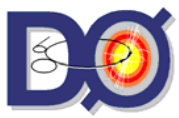
Grid Computing Centre Karlsruhe - The mission

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

Including resources for BaBar Tier-A

month/year	11/2001	4/2002	4/2003	4/2004	2005	2006	2007
CPU (kSI95)	**1**	**10**	**25**	**60**	**150**	**325**	**900**
disk (TByte)	**7**	**45**	**113**	**210**	**440**	**850**	**1500**
tape (Tbyte)	**7**	**111**	**211**	**350**	**800**	**2000**	**3700**

▲ starts in 2001 !

+ services .... + other sciences

Organization of GridKa

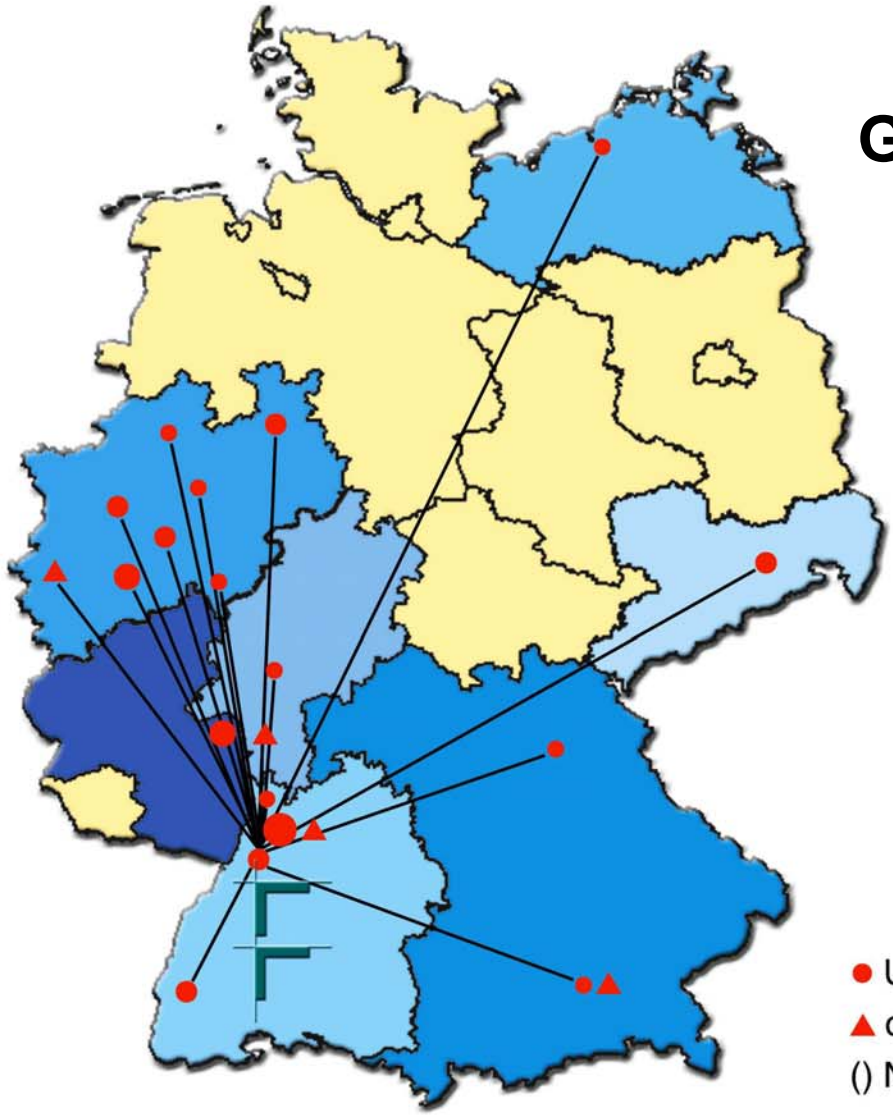
- **Project Leader & Deputy**
  H. Marten, M. Kunze
- **Overview Board**
  controls execution & financing, arbitrates in case of conflicts
- **Technical Advisory Board**
  defines technical requirements

It's a project with 41 user groups from 19 German institutions

German users of GridKa

- 19 institutions
- 41 user groups
- ~ 350 scientists

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

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

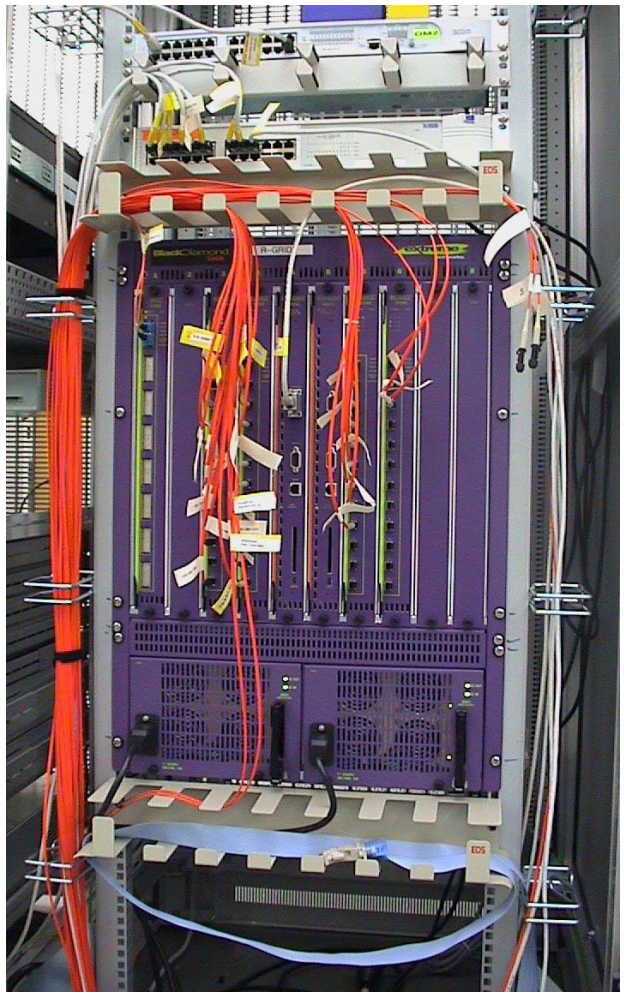
- 2 GB ECC RAM, 4x 80 GB IDE-Raid5, 2x Gbit Ethernet
- Linux & basic software on demand:
  RH 6.2, 7.1.1, 7.2, Fermi-Linux, SuSE 7.2

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

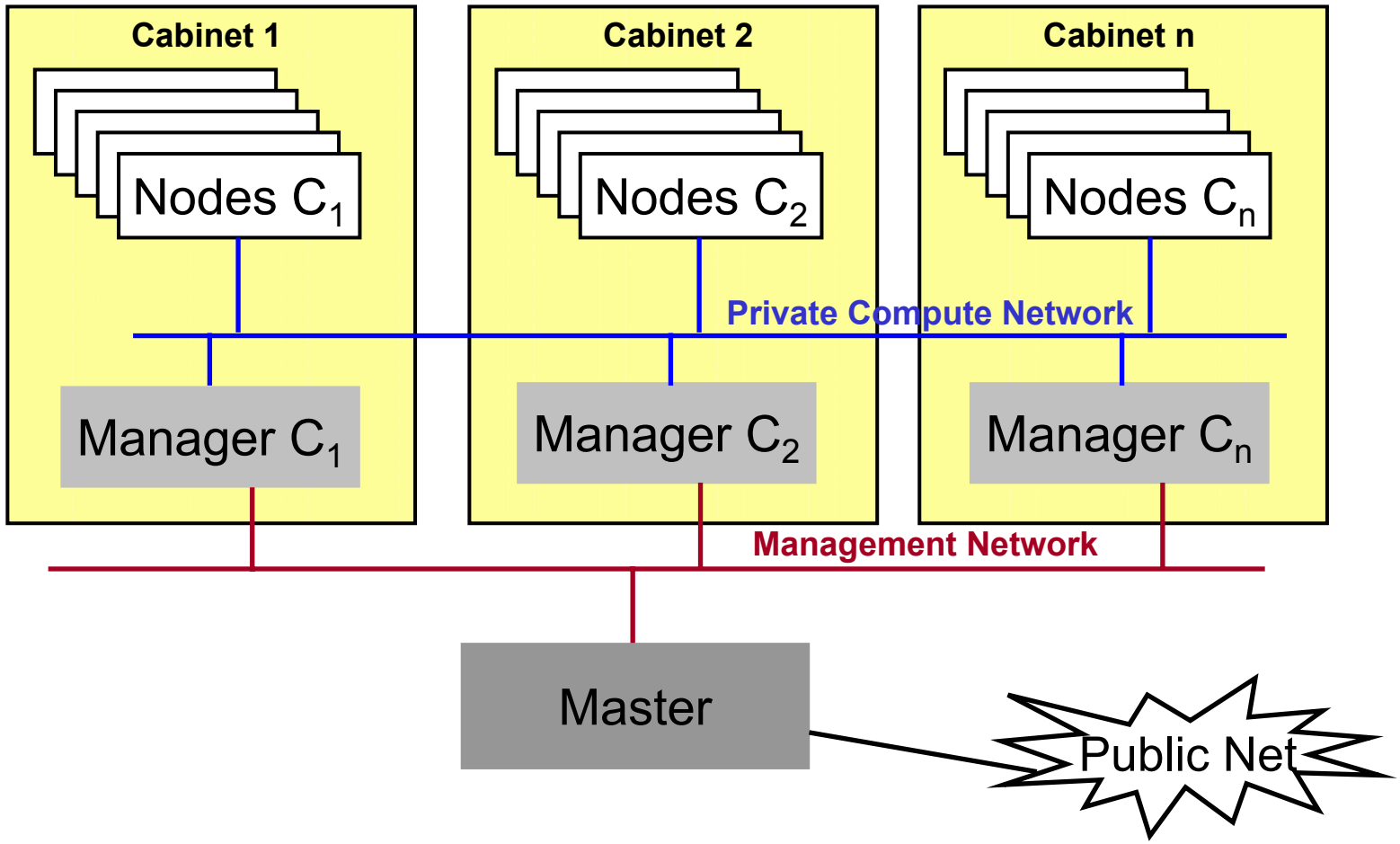
- scalability: many nodes to install and maintain (ca. 2000)
- heterogeneity: different (Intel-based?) hardware over time
- consistency: software must be consistent on all nodes
- manpower: administration by few persons only

➡ **This is for Administrators, not for a Grid Resource Broker**

Philosophies for Cluster Management

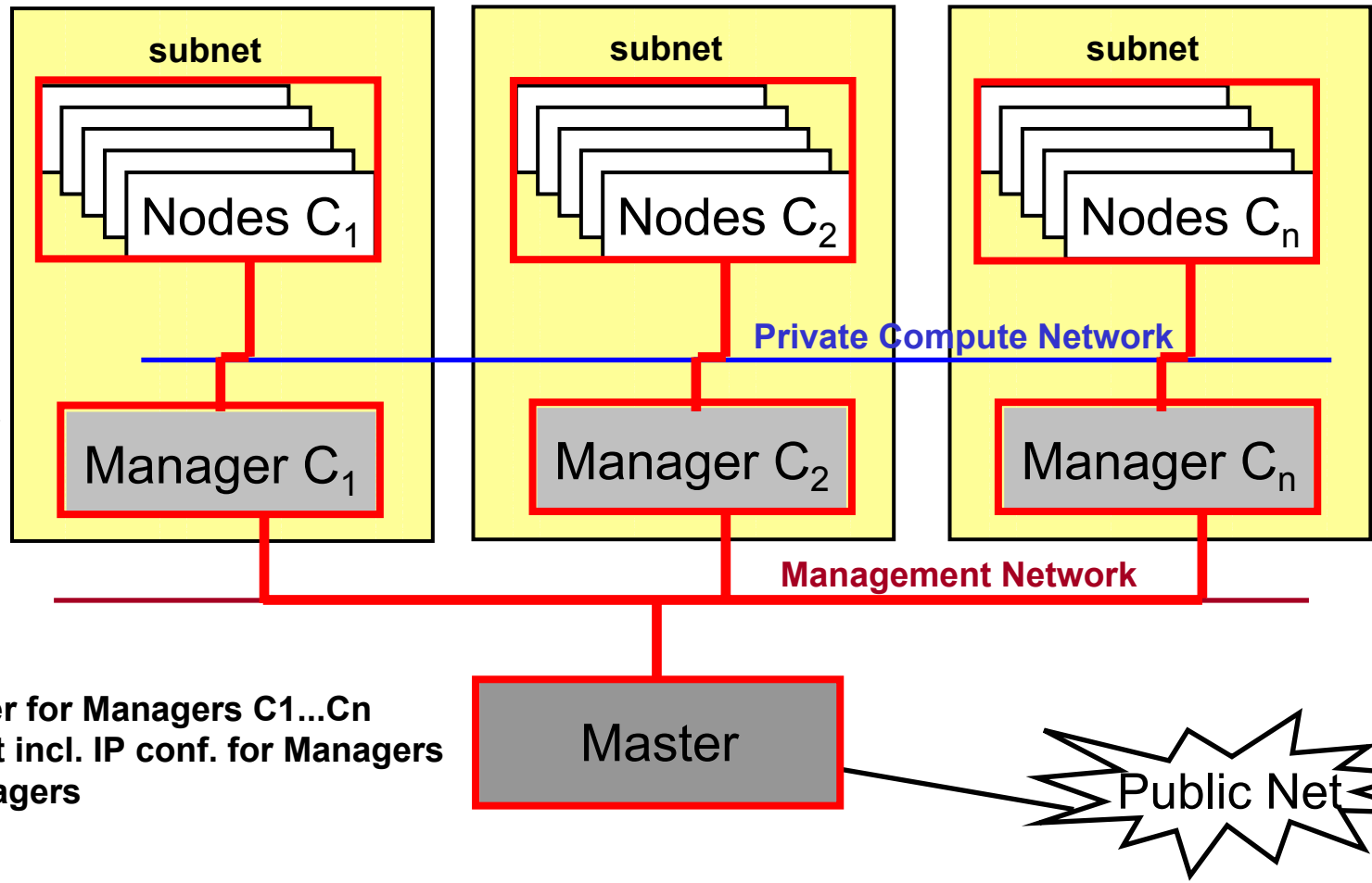
- scalability:
  - hierarchical instead of pure central management
  - combined push and pull for management information
  - info & event handling via separate management network
- heterogeneity: rpm instead of disk cloning
- consistency: distribute software from a central service
- manpower: automatise as much as you can

Architecture for Scalable Cluster Administration



Naming scheme: C02-001...064; F01-003,...

Installation - NPACI Rocks with FZK extensions



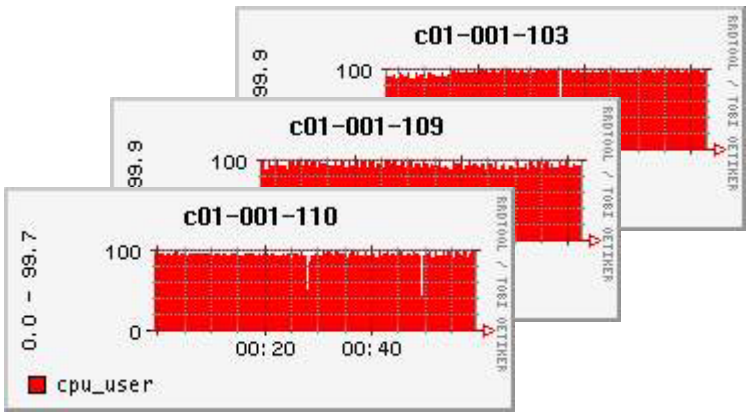
- **DHCP-server**
- **compute IP**
- **install nodes**

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

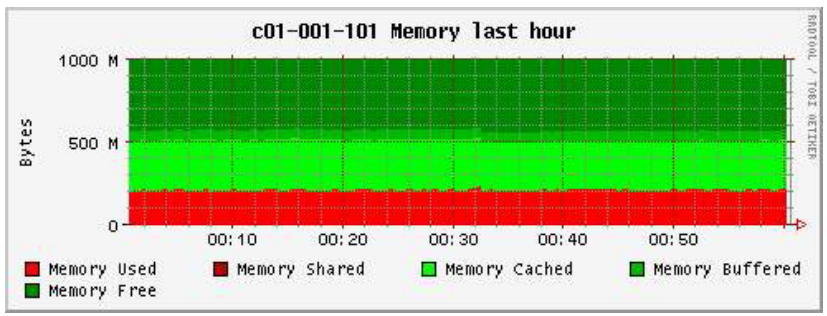
http://rocks.npaci.edu

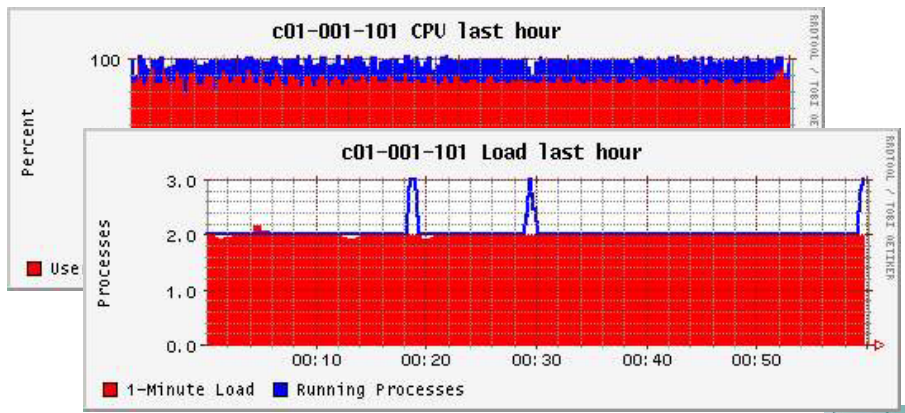
reinstall all nodes in < 1 h

System Monitoring with Ganglia



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

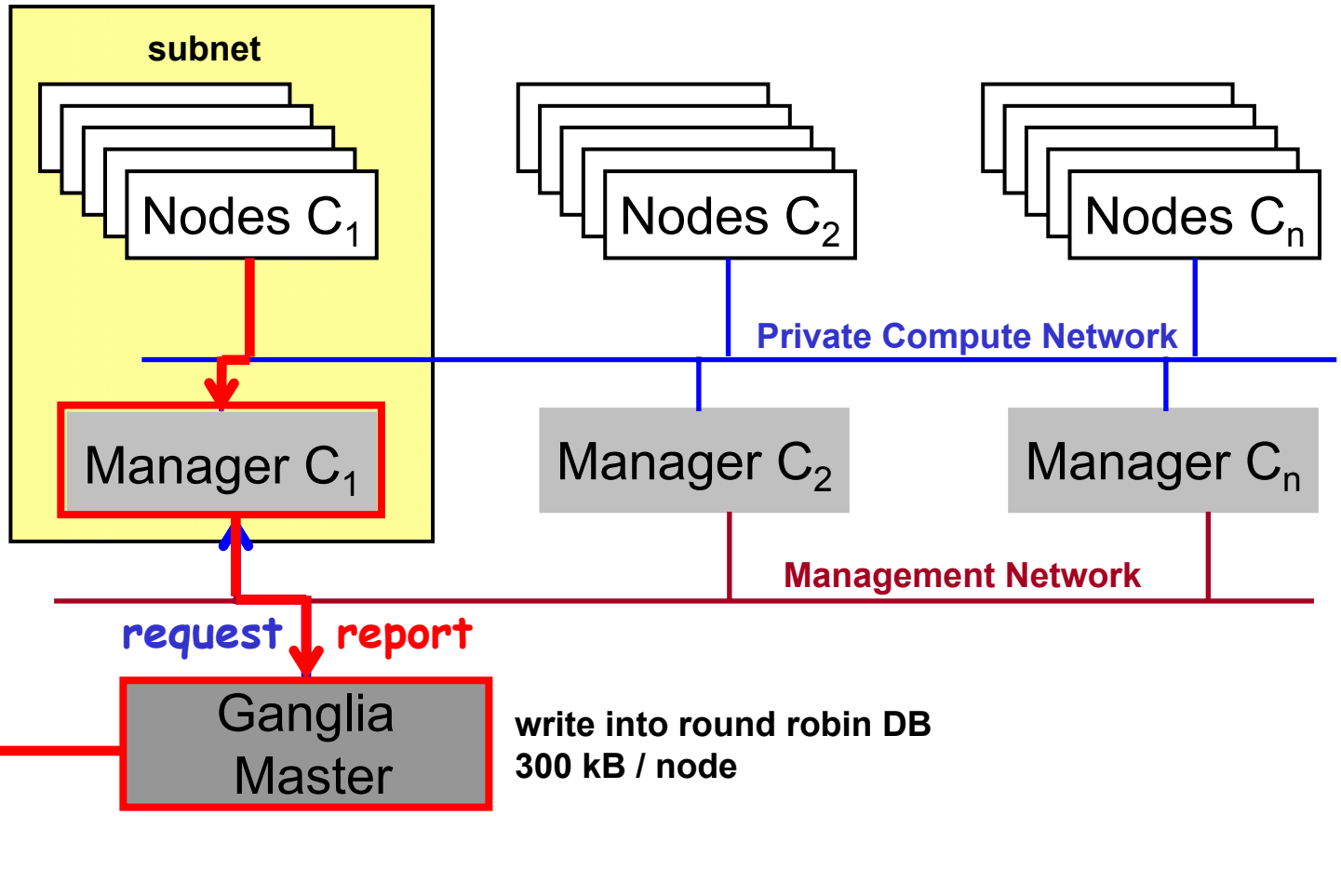




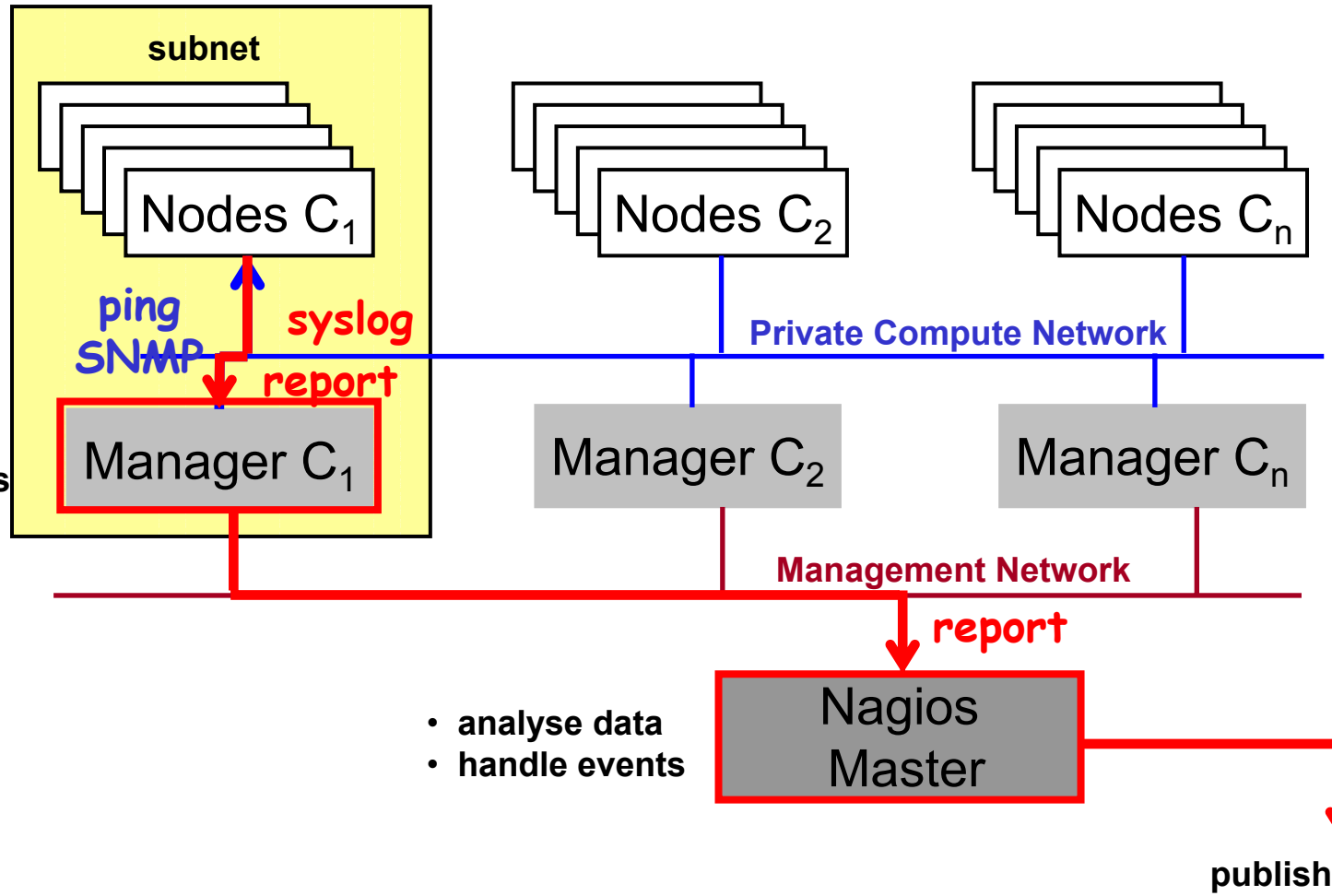
http://ganglia.sourceforge.net

System Monitoring with Ganglia - Combined Push-Pull

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



Cluster Management with Nagios - Combined Push-Pull



- analyse data
- handle local events

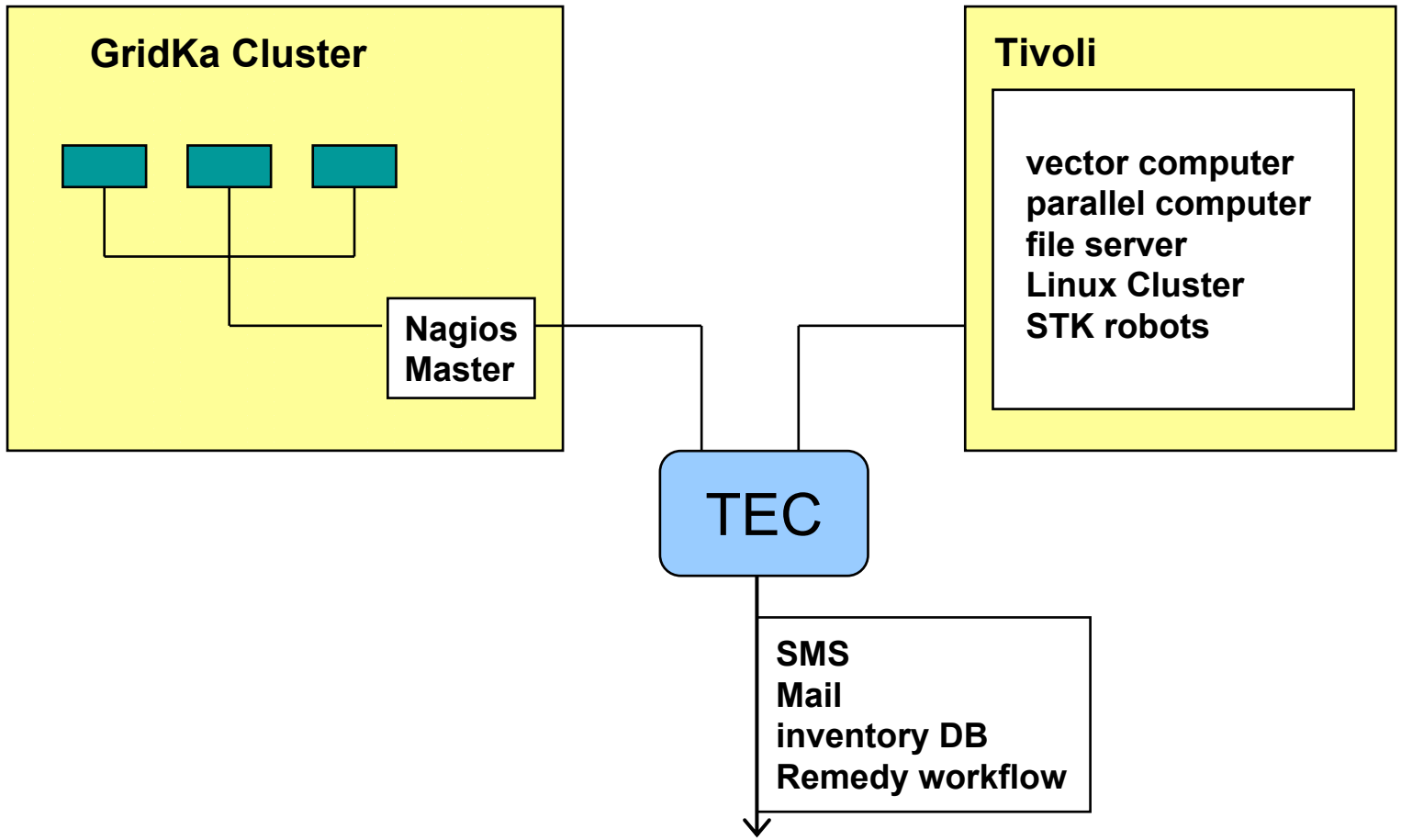
- analyse data
- handle events

GPL

http://www.nagios.org

GridKa

Combining Nagios with Tivoli Enterprise Console ?

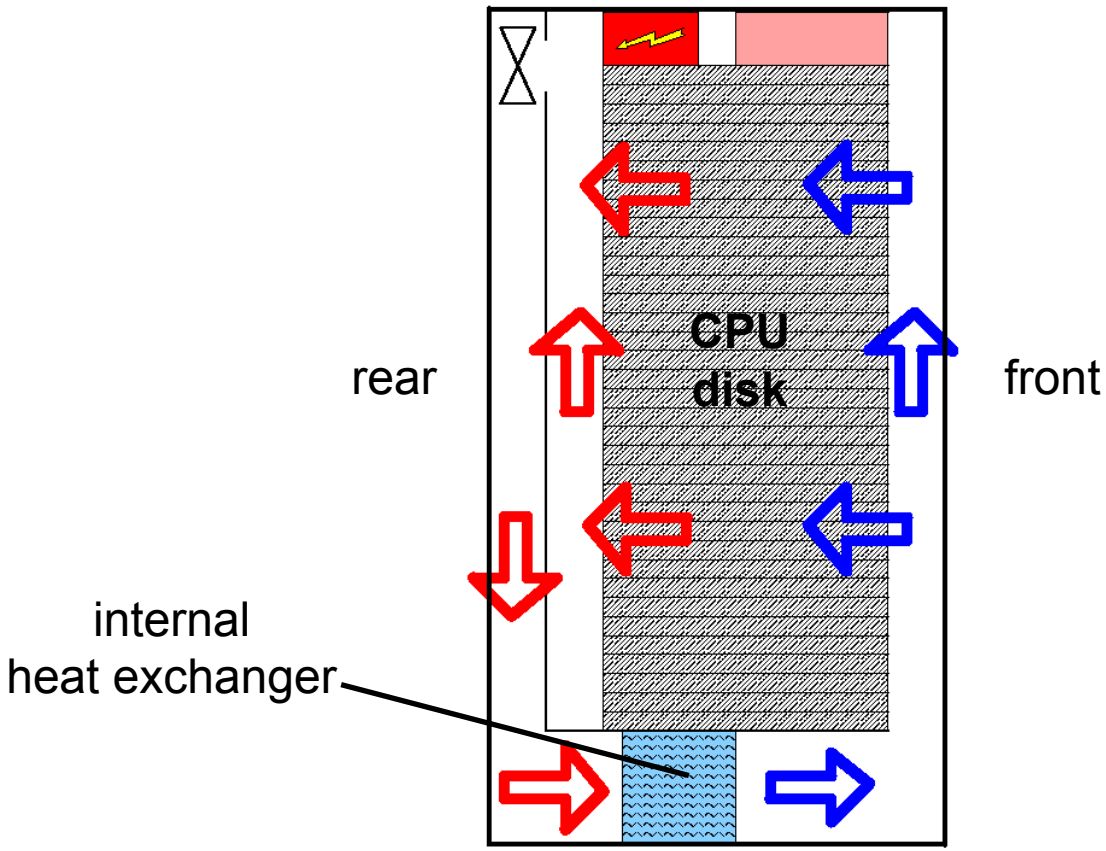


Infrastructure

We all want to build a Linux cluster...

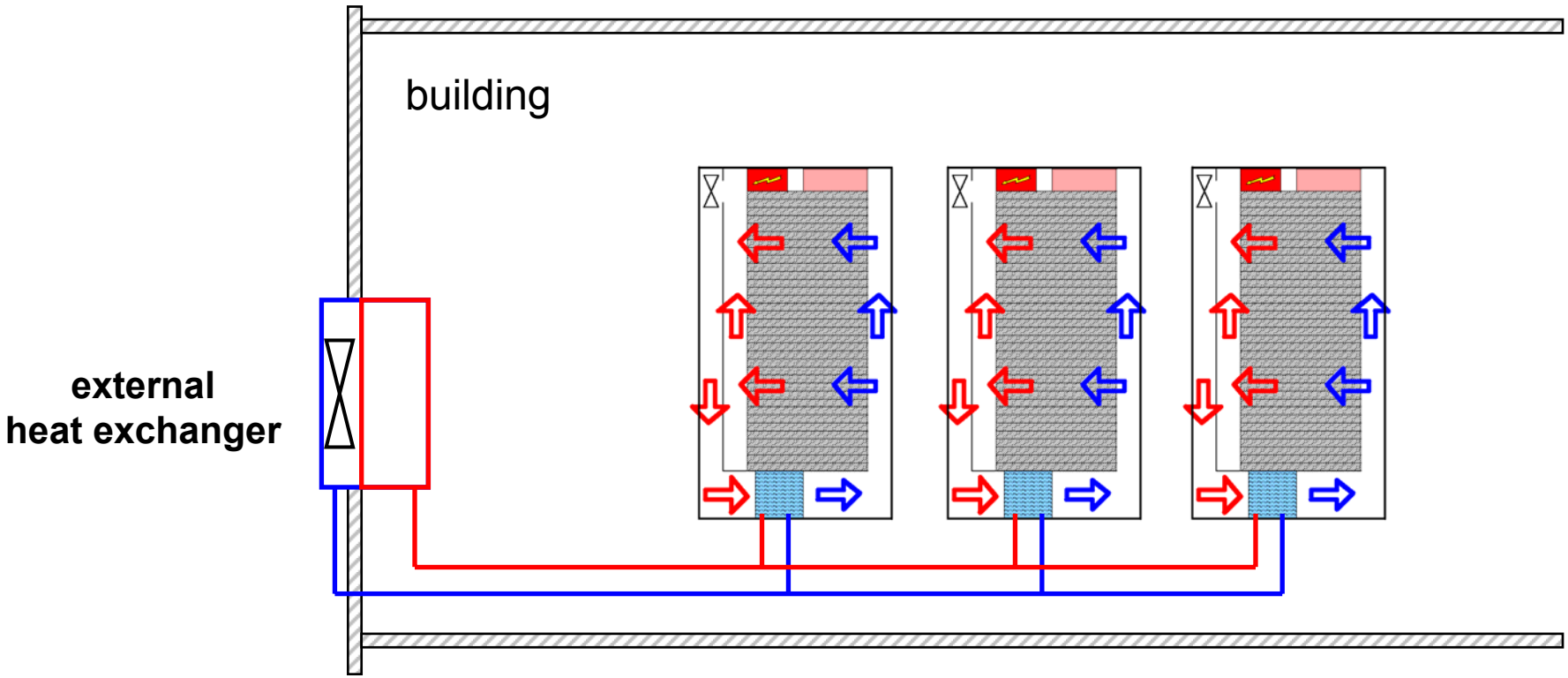
... do we have the cooling capacity?

Closed rack-based cooling system - A common development of FZK and Knürr

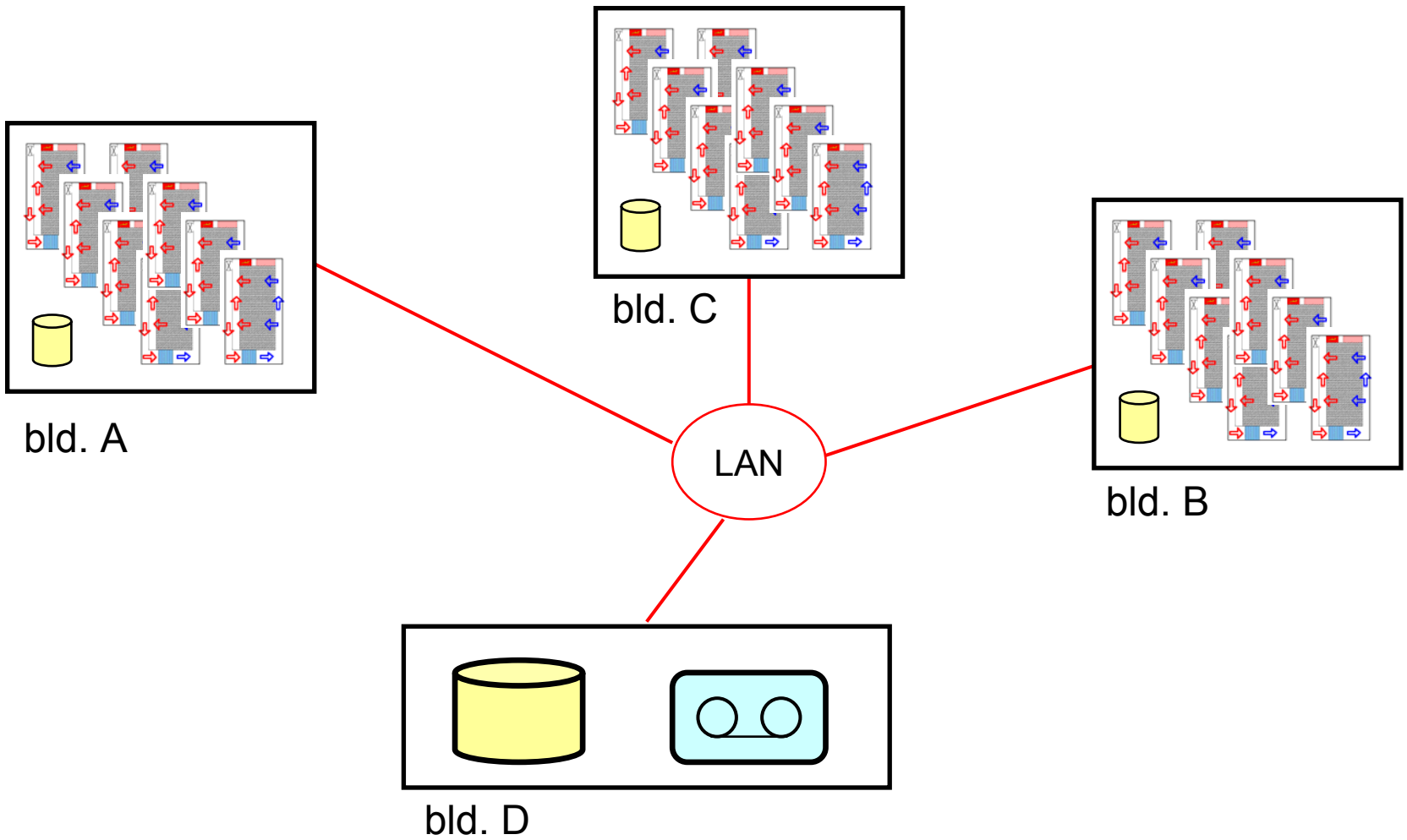


- 19” technique
- 38 units height usable
- 70x 120 cm floor space
- 10 kW cooling
- redundant DC fans
- temperature controlled CPU shut-down

Closed rack-based cooling system -
Estimated cost reduction > 70% compared to air conditioning



Build a cluster of clusters (or a Campus Grid ?)

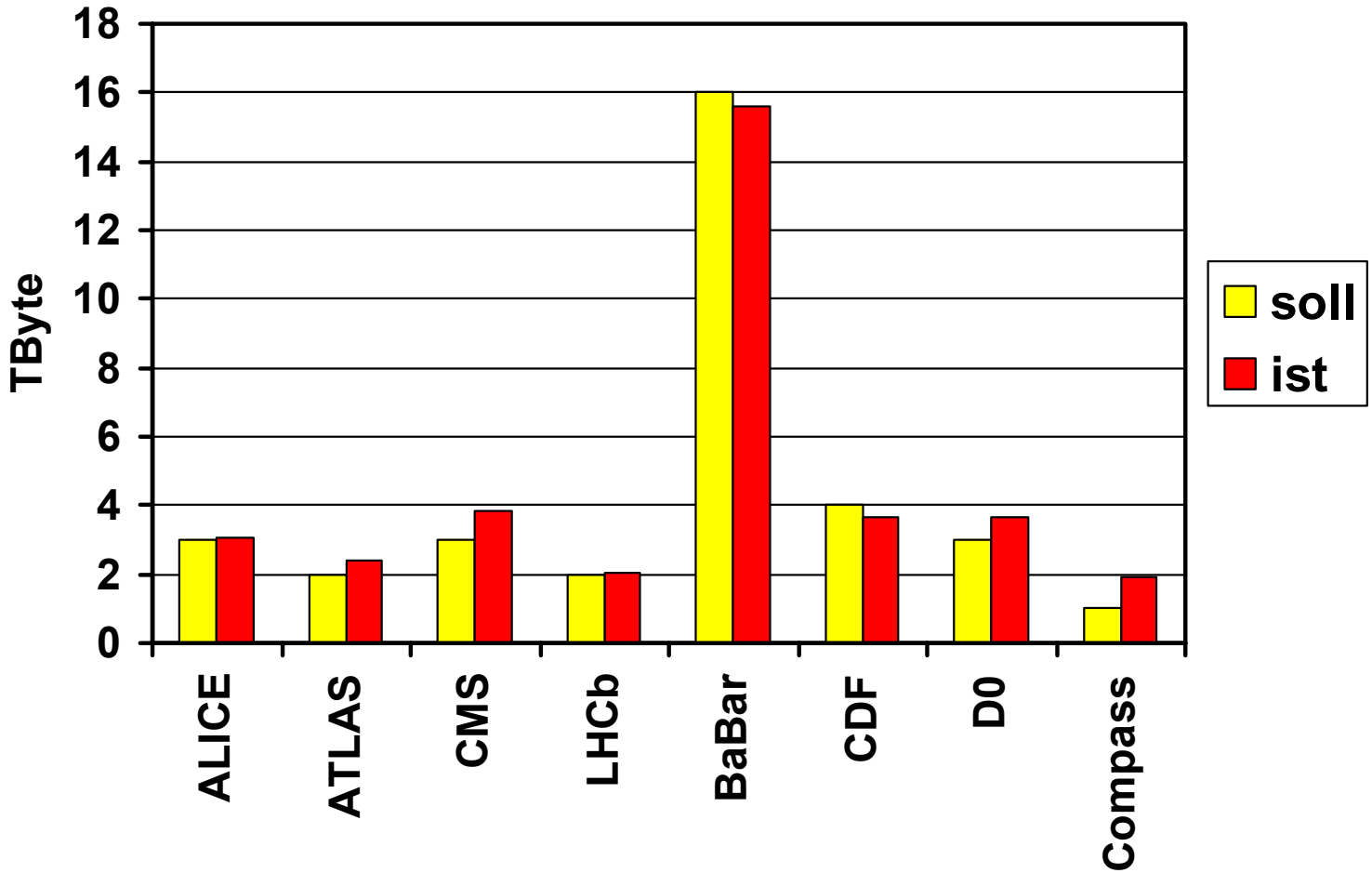


Online Storage

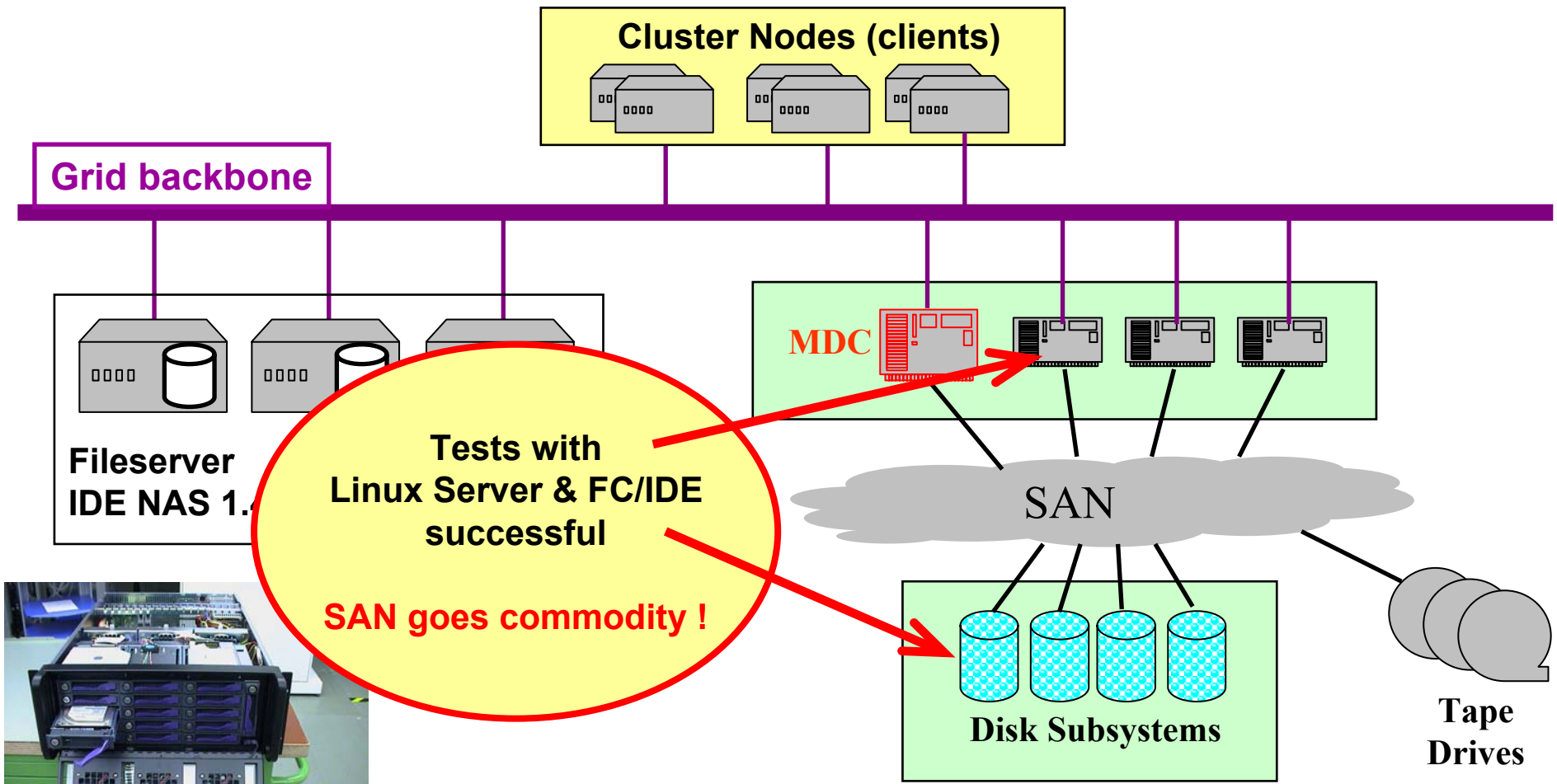
- 59 TB brutto
- 45 TB net capacity
- ~ 500 disk drives
- mixed IDE, SCSI, FC
- ext3 & ufs file systems

- **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



Scheme of (Disk) Storage



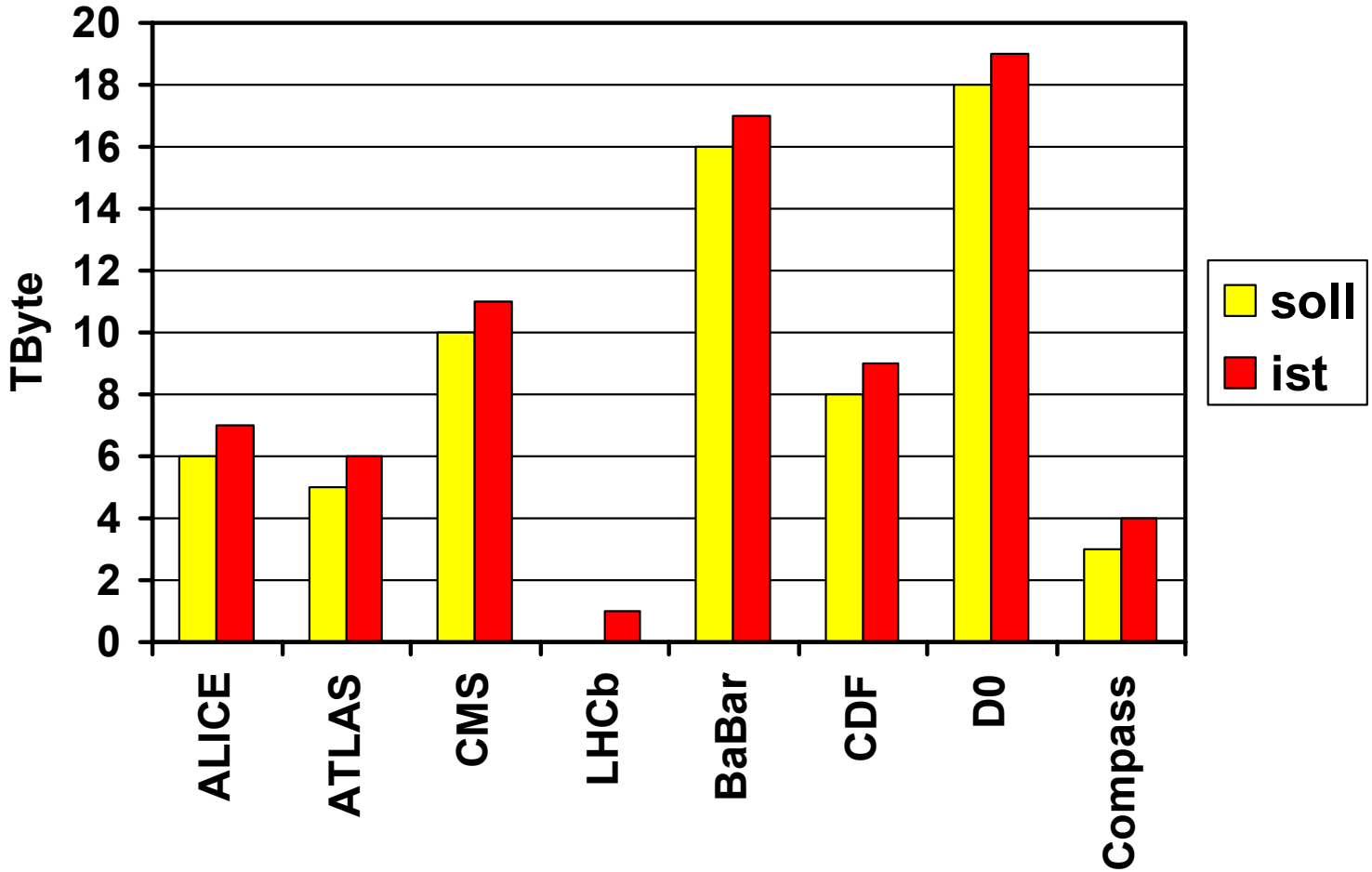
Disk Storage & Management – does it scale?

- **>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 <u>now</u> !**

We will test gfs & gpfs for Linux

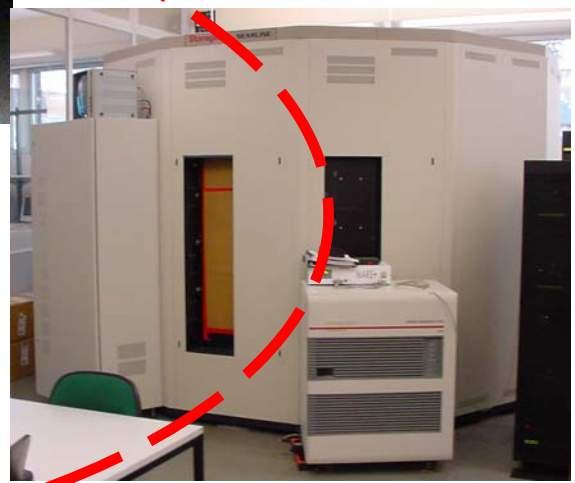
Available Tape Space for HEP: 106 TByte native



GridKa Tape System

FZK Tape 1

FZK Tape 2



FZK SAN

IBM 3584

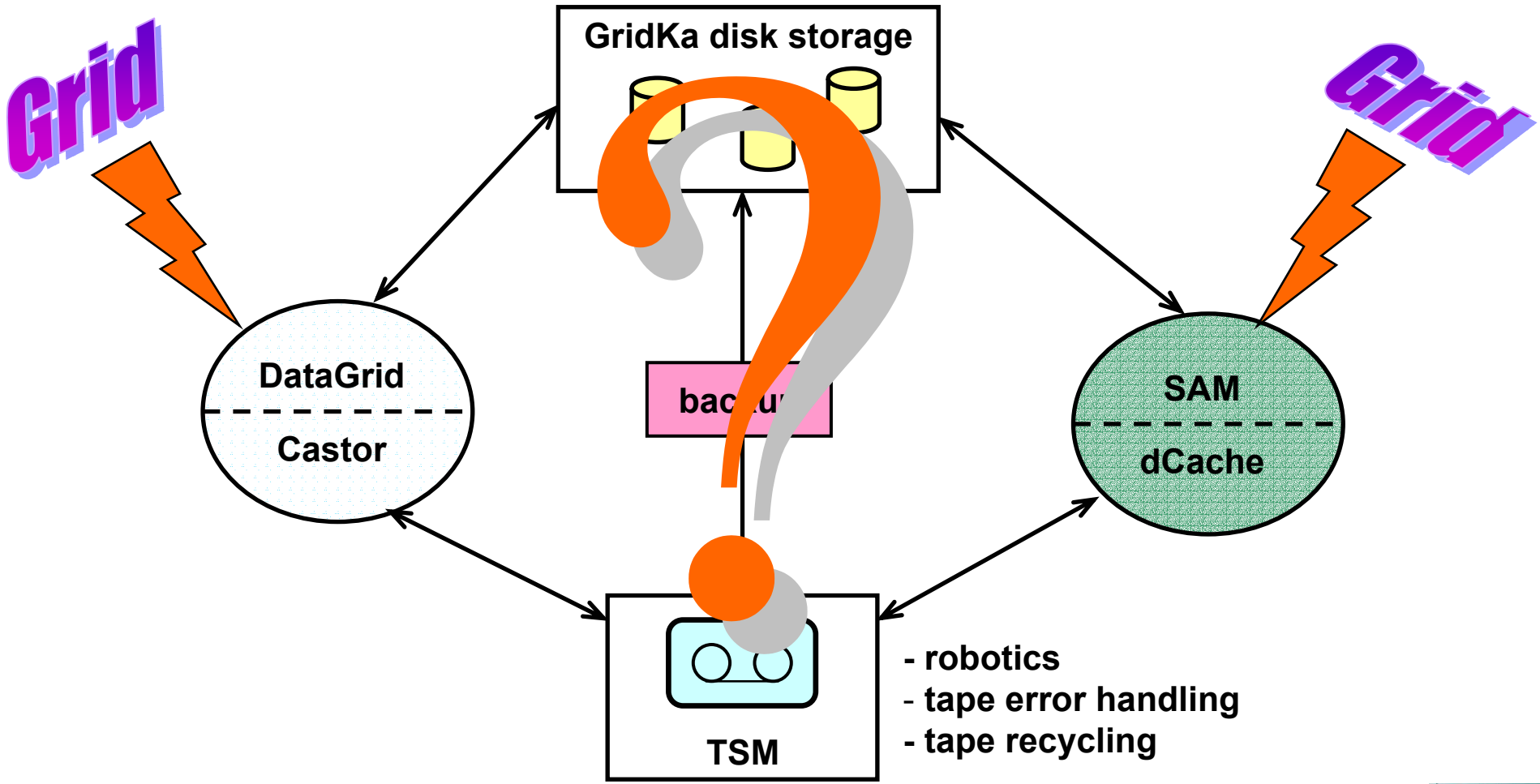
- ~ 2400 slots LTO Ultrium
- 100 GB/tape
- 100 TB native available
- 8 drives, 15 MByte/s each
- Backup/Archive with Tivoli Storage Manager

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 -- DataGRID **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.