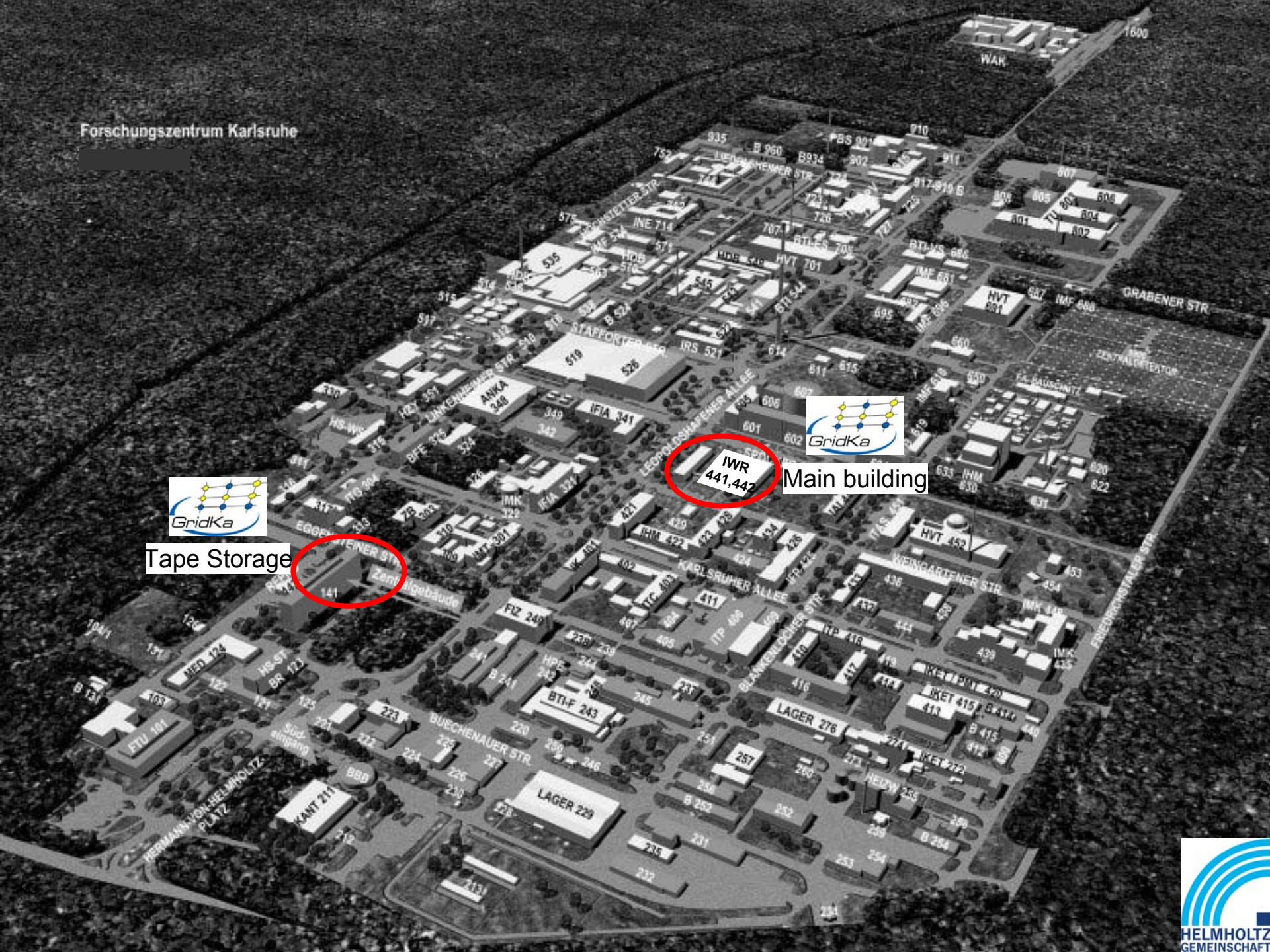
Status and Plans for the Tier1 Karlsruhe (D)

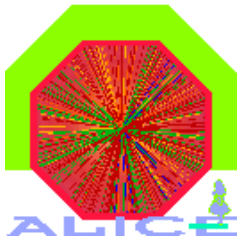
Forschungszentrum Karlsruhe GmbH
Institut für Wissenschaftliches Rechnen
Postfach 3640
D-76021 Karlsruhe

Holger Marten, Axel Jäger,
Bernhard Verstege, Jos van Wezel

http://grid.fzk.de


Forschungszentrum Karlsruhe
GridKa
Tape Storage
IWR
441,442
Main building
GridKa
HELMHOLTZ
GEMEINSCHAFT

High Energy Physics Experiments served by GridKa



LHC experiments

(SLAC, USA)

(FN

- Committed to Grid Computing
- Have real data already today

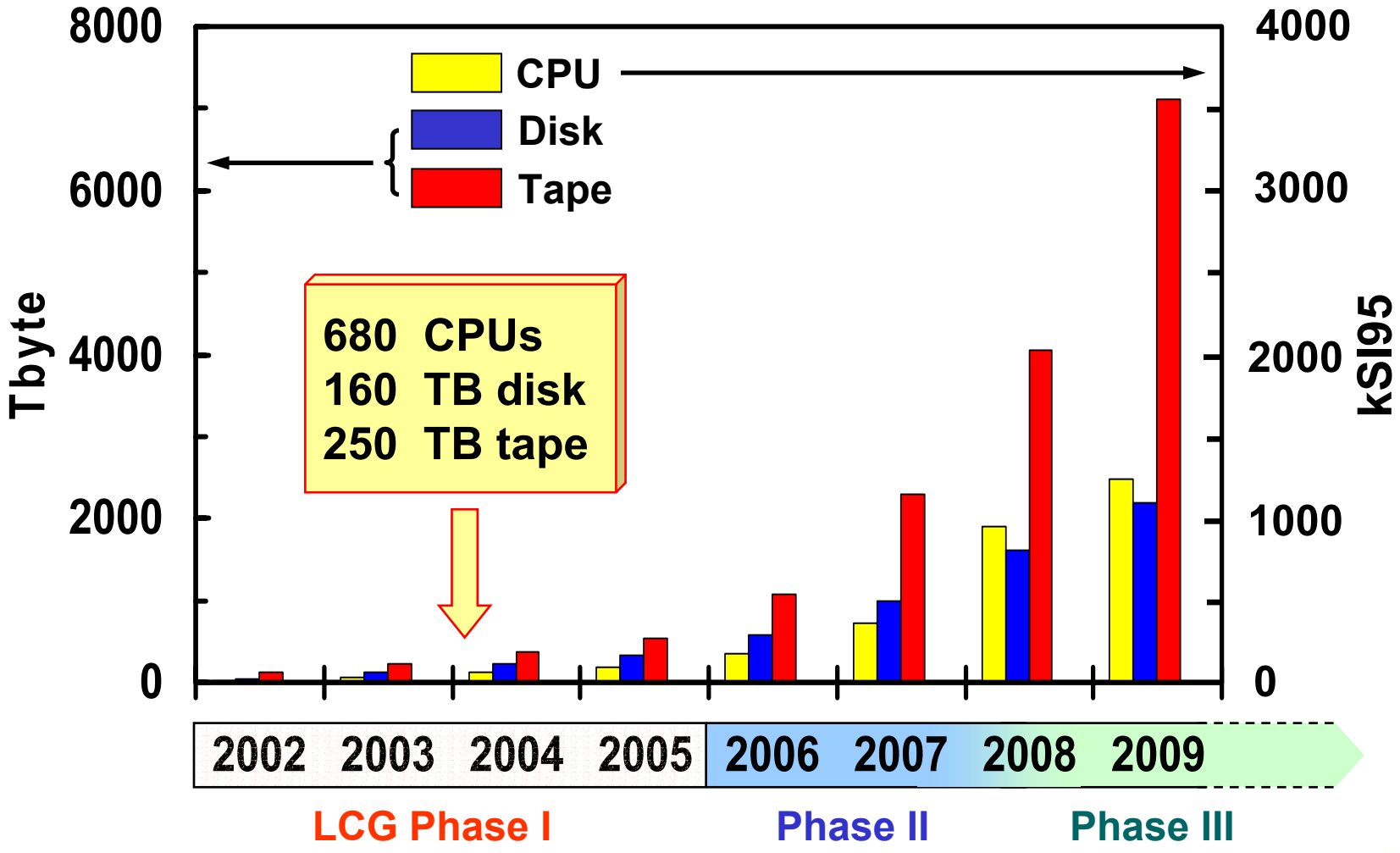
(FNAL ,USA)

(CERN)

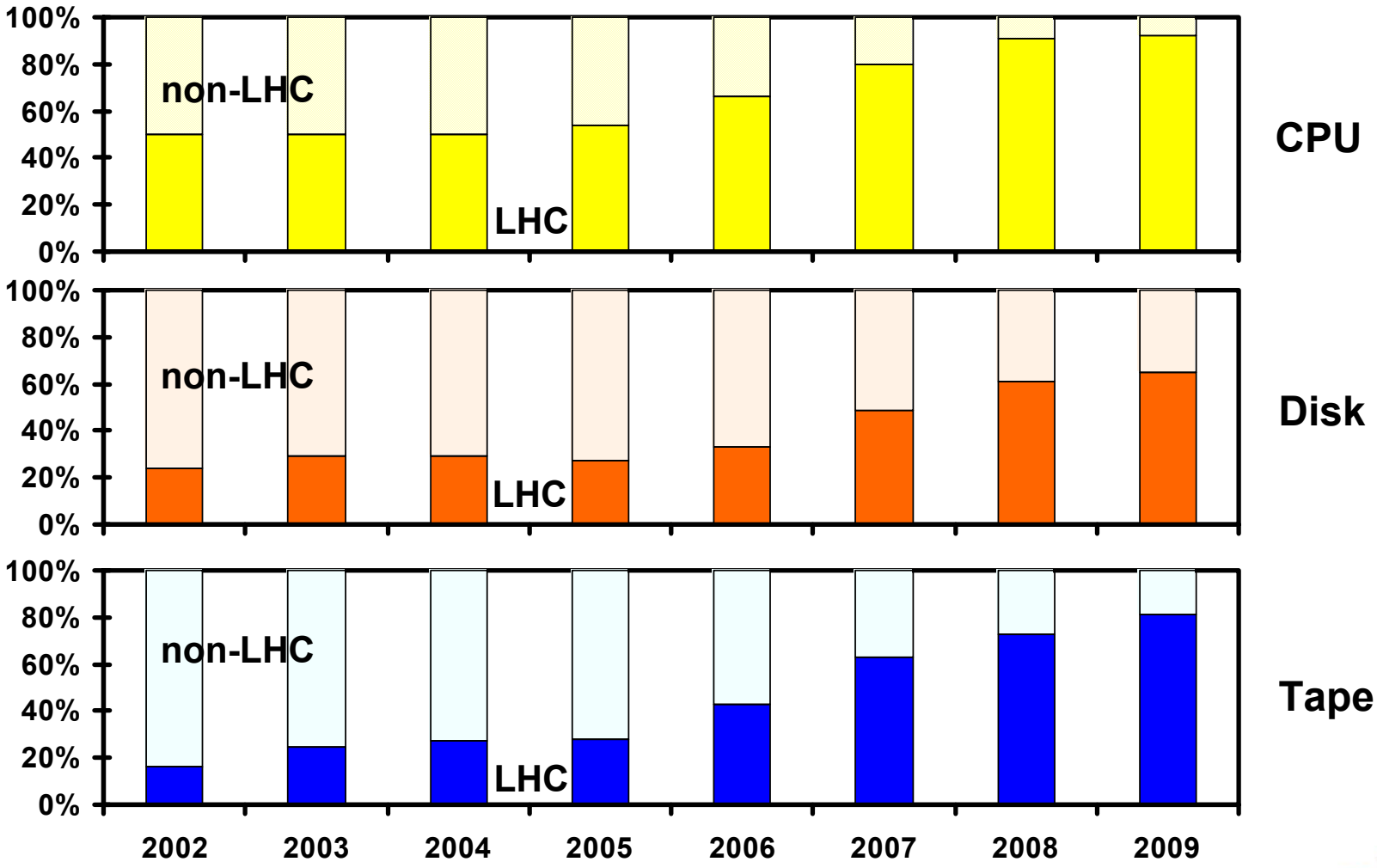
non-LHC experiments

Other sciences later

GridKa planned resources



Distribution of planned resources at GridKa



GridKa Installation and Plans

Address Bernd Panzer-Steindels topics:

- infrastructure (electricity, cooling)
- networks (LAN / WAN)
- worker nodes (batch, installation & management)
- purchase quality assurance
- storage
- LCG / EGEE

Infrastructure

Use existing infrastructure of FZK (buildings, electricity, ...)
Special Technical Infrastructure Division at FZK

Floor space

- ~600 m^2 floor space reserved for GridKa

Electricity

- 20kV supply network
- several transformers with 2 MW for computing centre available

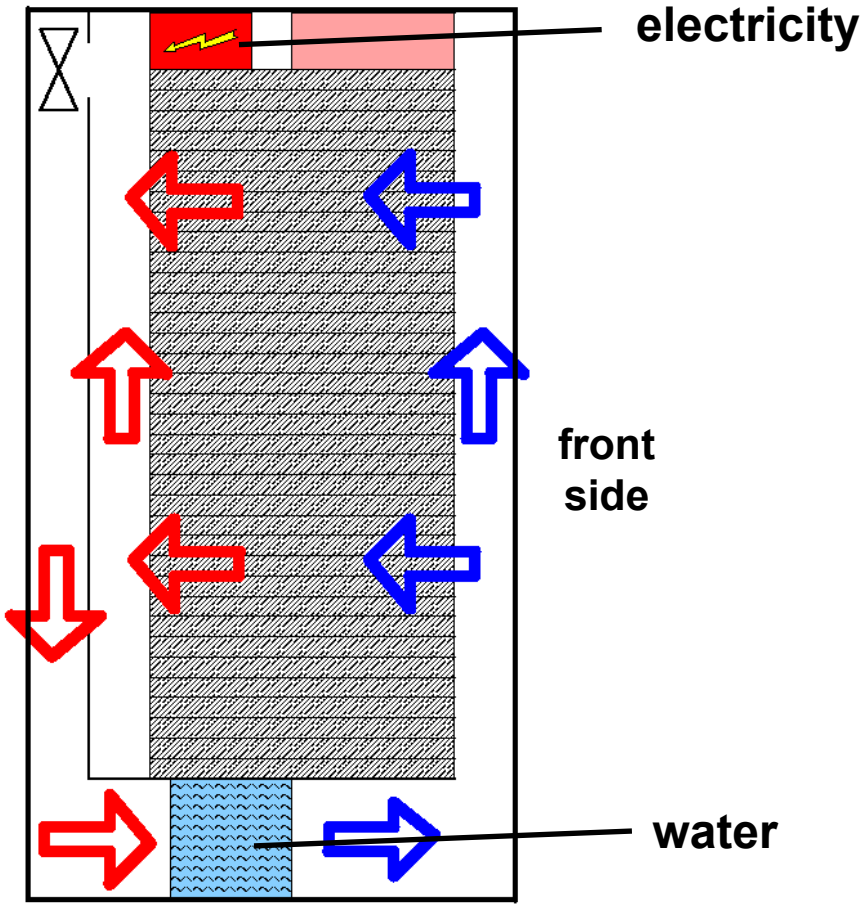
Cooling

- 3 external heat exchangers with 1.5 MW available
- however, air condition was an issue !
  (reduced cross sections of air flow channels)

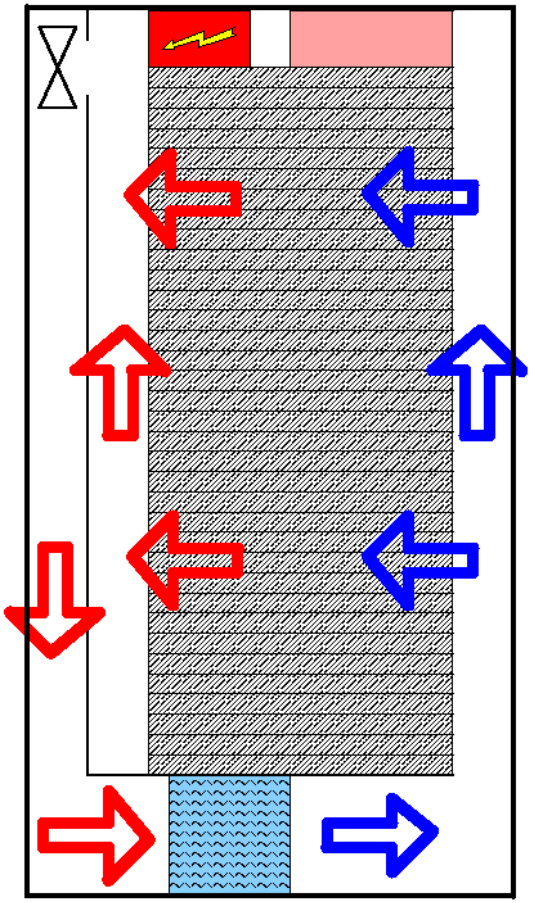
Power rails at the ceiling

- separate well defined rails for “normal” power and USV
- prevents from (un)plugging wrong cables
- prevents from electrical hazards
- separates electricity from water

Equipment cabinet with water cooling



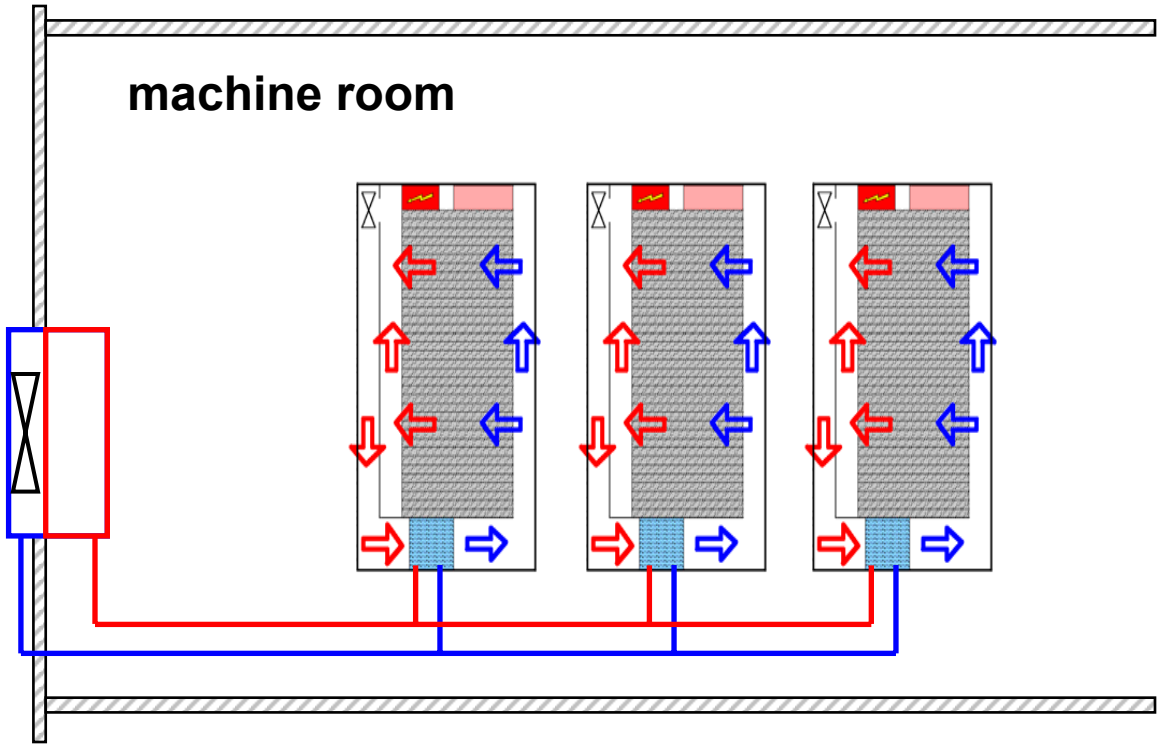
Equipment cabinet with water cooling



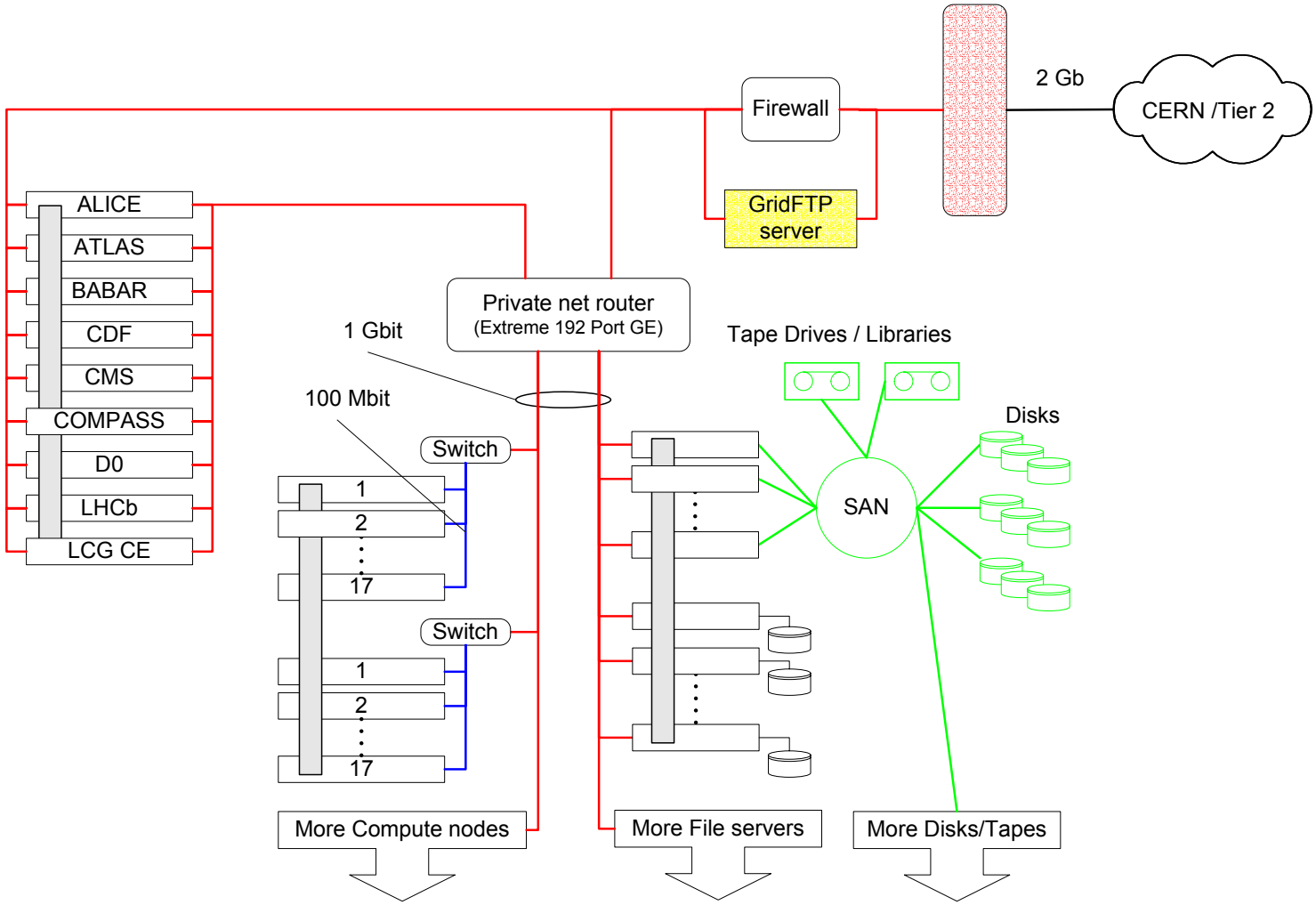
- 19'' technique
- 37 units usable height
- 70x 120 cm floor space
- 10 kW cooling
- redundant DC fans
- temperature regulated fans
- temperature controlled
  - 22° warning
  - 26° critical
  - 30° power off
- internal smoke detector
- SNMP monitored
- manual reset after power off !

Cooling system – overall design
no air condition needed

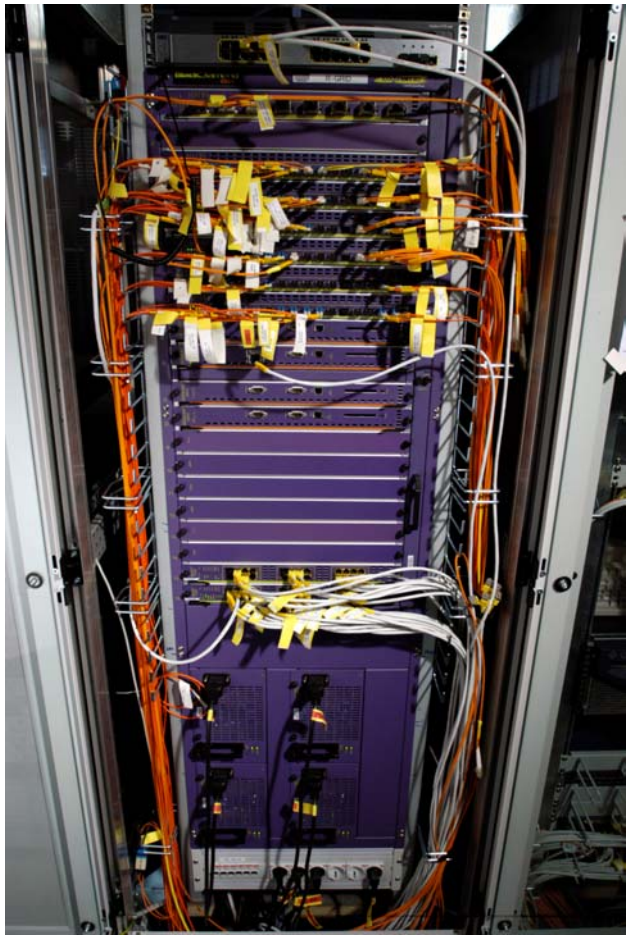
Capacity
- 200 KW today
- 1000 KW Dec 2004



Network diagram



Network components



Extreme Black Diamond 6816

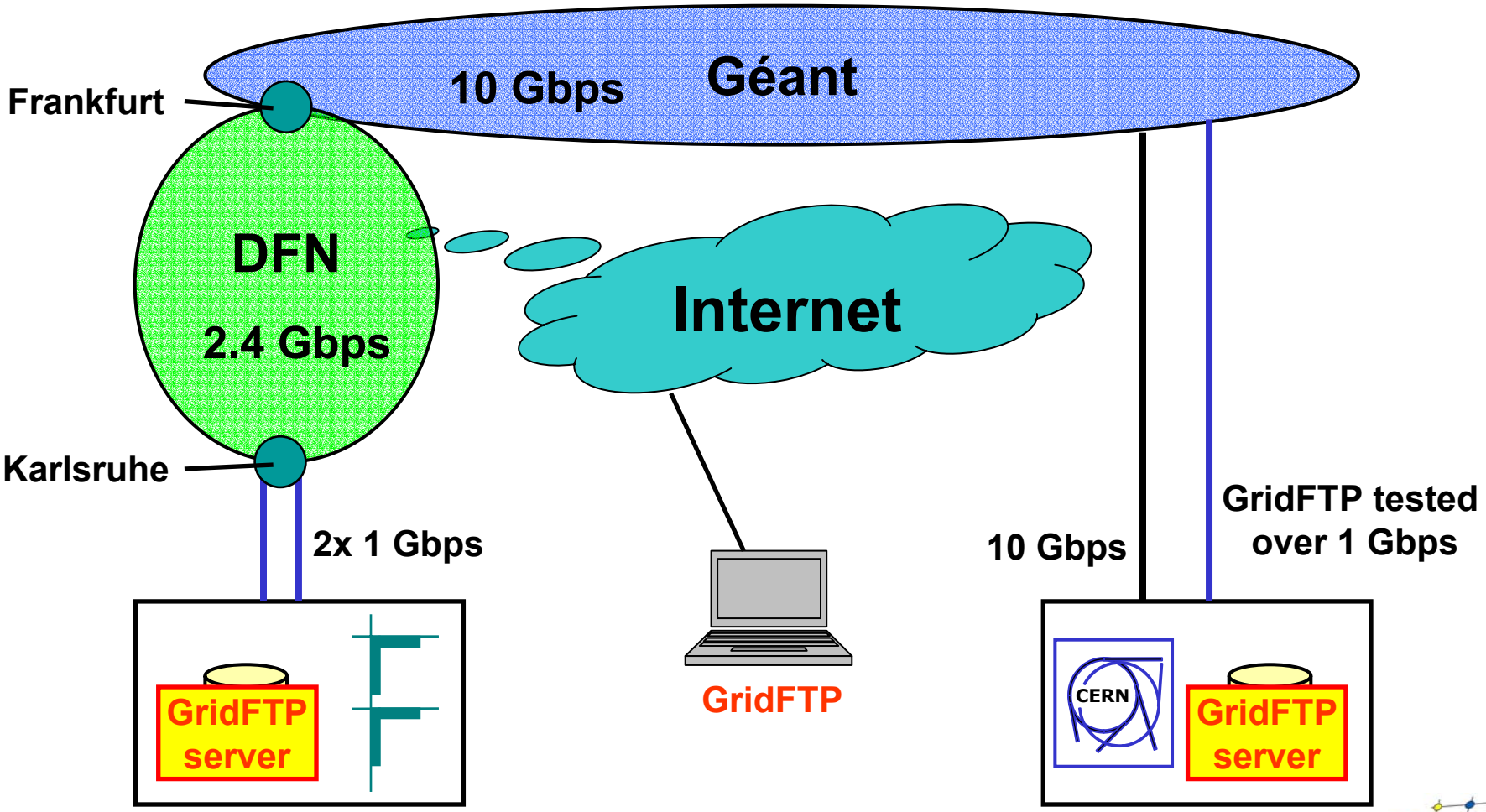
- 256 Gbps backplane
- max. 192 Gbit ports
- redundant power supplies
- redundant management boards

20 Extreme 2816 Switches

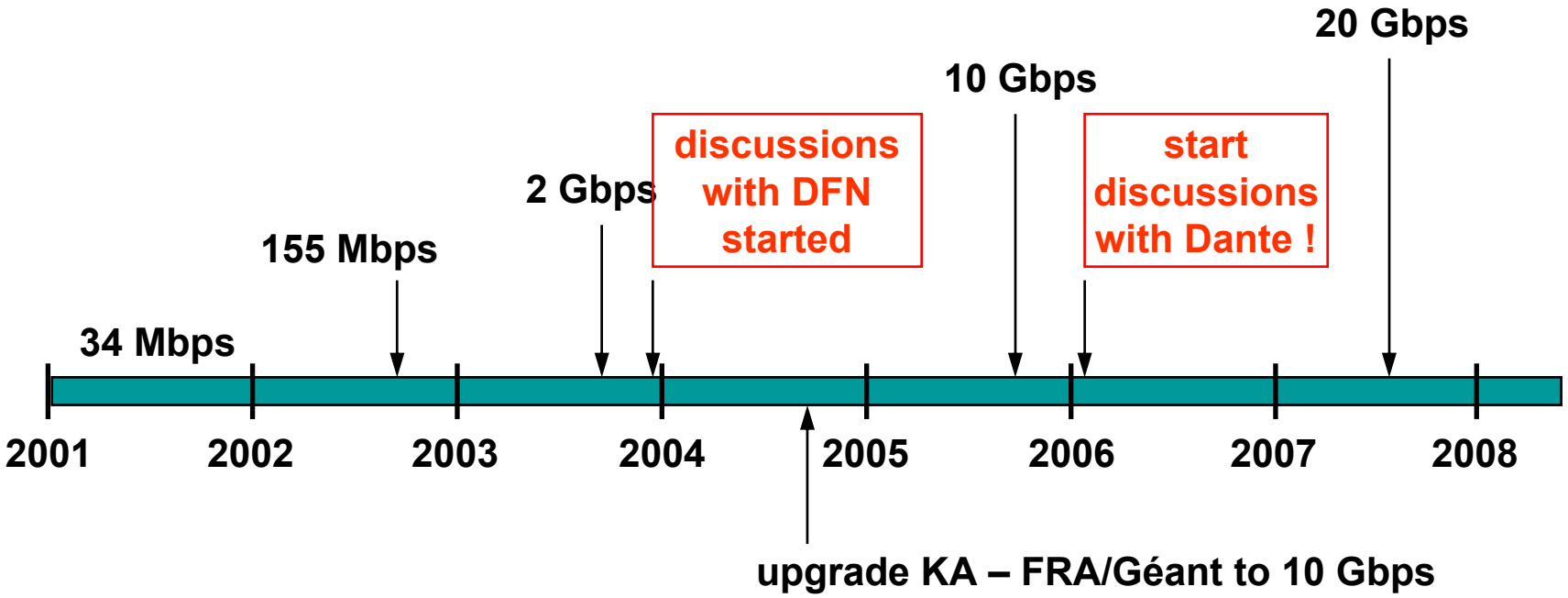
- 2 Port GE 24 Port FE

Cisco PIX Firewall

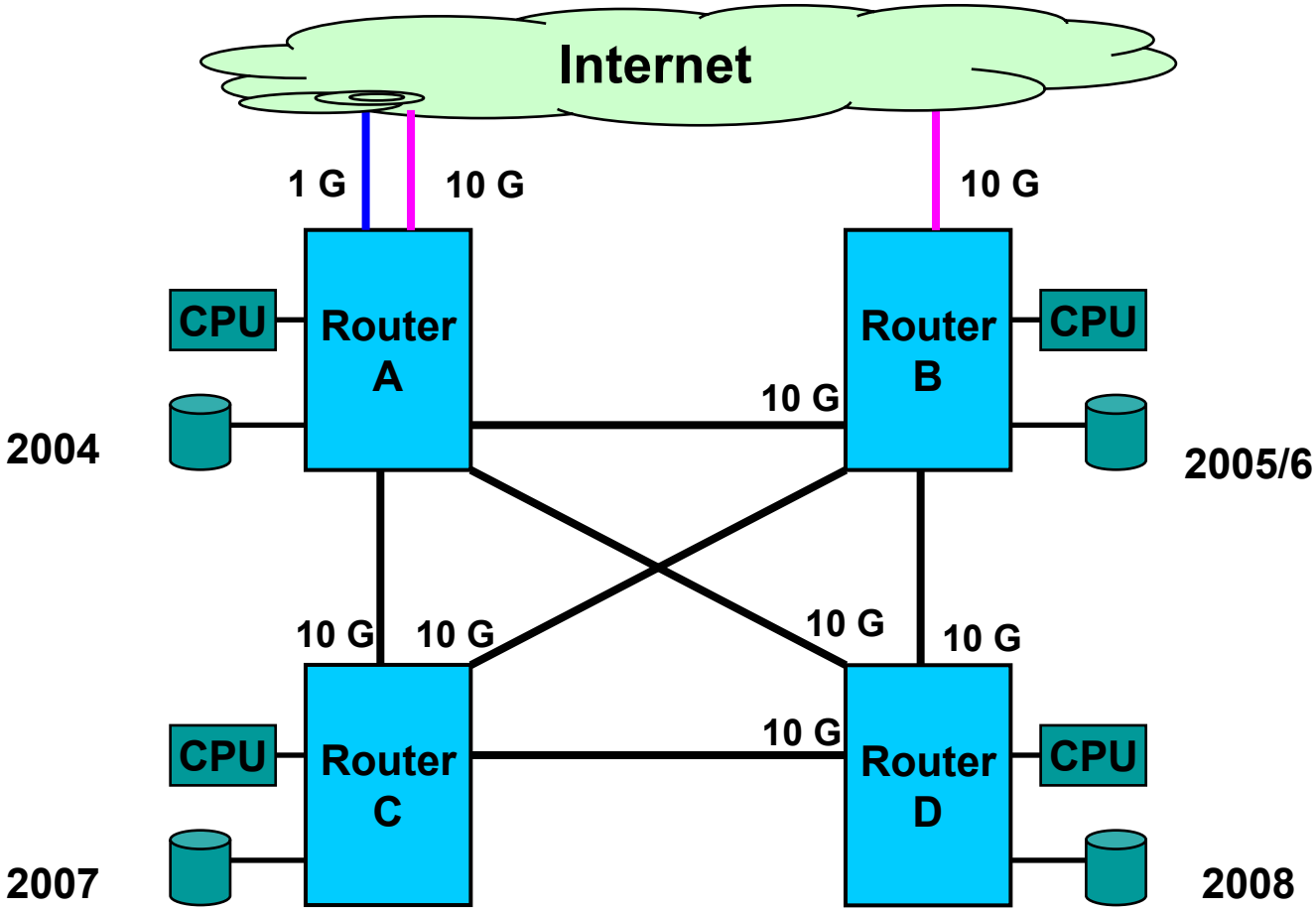
WAN connectivity and Gigabit test with CERN



Planning for WAN connectivity



Plan for LAN/WAN extensions under discussion



Worker Nodes & Testbeds

Production environment

97x	dual PIII, 1,26GHz	97 kSI2000	
64x	dual PIV, 2,2 GHz	102 kSI2000	
72x	dual PIV, 2,667 GHz	130 kSI2000	
108x	dual PIV, 3,06 GHz	216 kSI2000	(72 used for LCG-2)
53x	dual PIV, 3,06 GHz	106 kSI2000	ordered for July
140x	dual PIV ??	270 kSI2000	to be ordered for October

Test environment

- additional 30 machines in 5 test beds with EDG 1.4.x, LCG-1, LCG2

Issues

- heterogeneous hardware from many different vendors
- old OS (RH 7.3) doesn't always support new hardware
- applications & middleware not certified for new OS or hardware (Opterons?)
- install new (Grid) software without disturbing the production
- number of test beds increases very fast (is EGEE the next?)

Batch system

GridKa switched from OpenPBS to PBSPro last November

Advantage of PBSPro vs OpenPBS

- clients contact server for status updates (better scalability, no pending/hanging server when client(s) die)
- possibility to read current fair share values
- fair share values can be applied to other resources then CPU time
- possibility to add local “resources” (example: Babar skimprod)
- plugin compatible with OpenPBS (users see no difference, extends existing configuration)

Disadvantage

- licence cost

We’ll discuss BQS with IN2P3 next week.

Purchase quality assurance

Not really an issue

- 3 years warranty for all equipment
- problem with small firm offers (Are they still there in three years?)
- use better/more reliable hardware for core services (batch server, file servers etc.)

Nevertheless

- call for tenders is tedious and time consuming
- always risks for failures or delays e.g. due to objections by potential suppliers

Installation and management

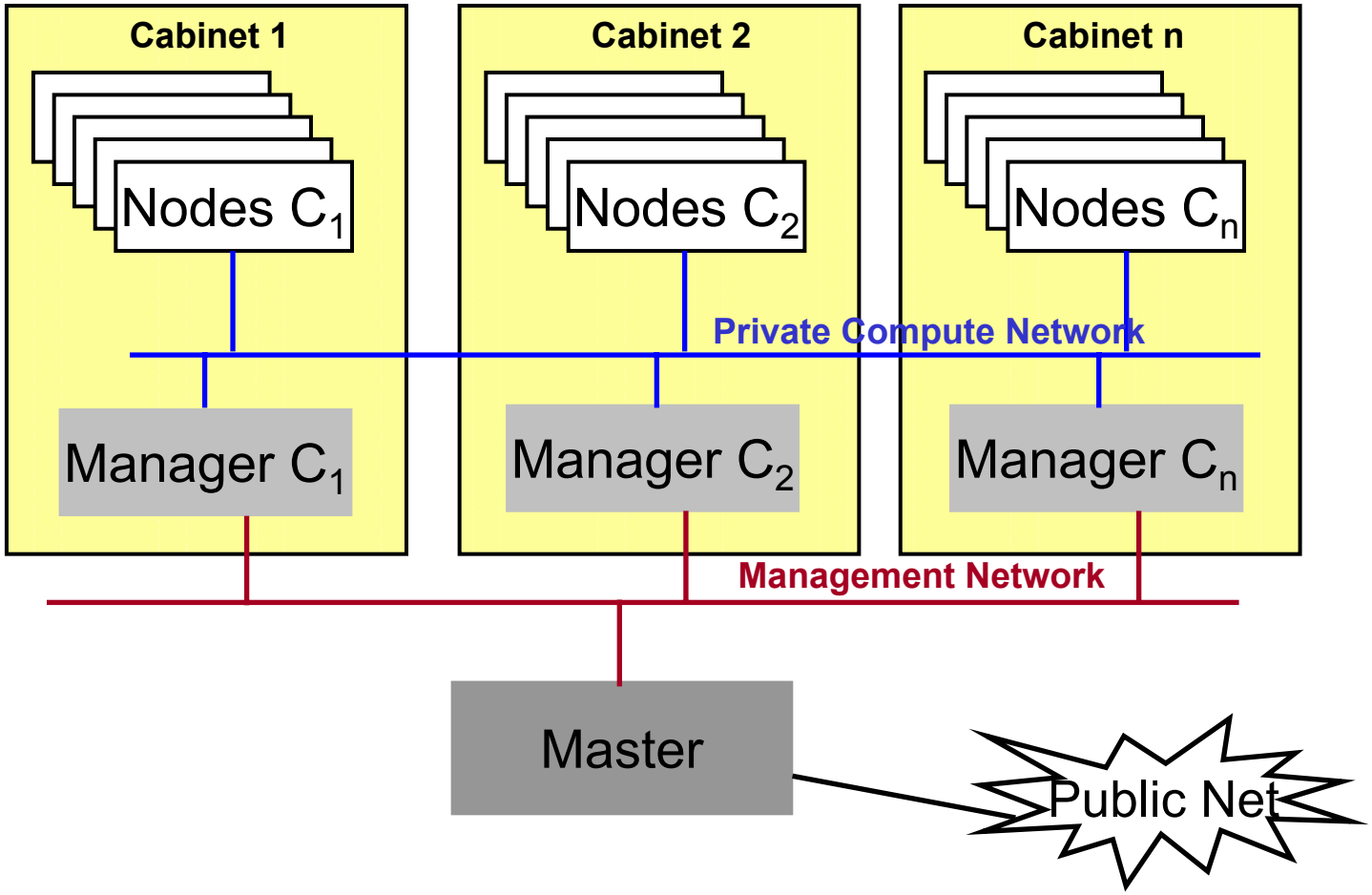
Software Installation
- NPACI Rocks with own extensions
- largest site working with NPACI Rocks **http://www.rocksclusters.org**

System Monitoring
- cluster usage for users available on the web
- Ganglia Cluster Toolkit **http://ganglia.sourceforge.net**

System Management
- "health monitoring" with NAGIOS **http://www.nagios.org**

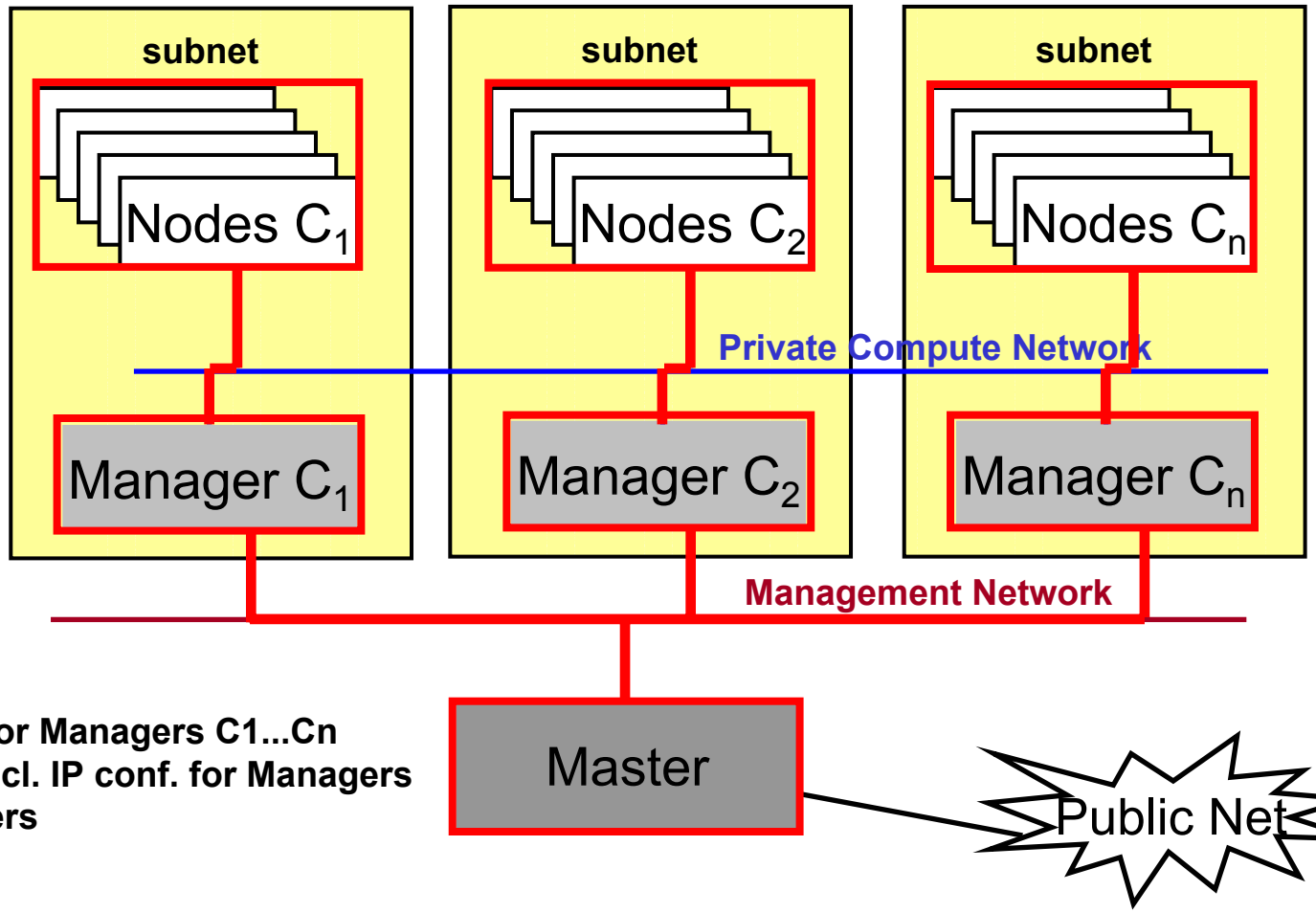
User statistics
- home made based on PBS logs

Architecture for Scalable Cluster Administration



Naming convention: C01-006-109, F01-002-003, ...

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://www.rocksclusters.org

reinstall all nodes in < 1 h

Storage

GridKa

- online data stored in NAS (40 TB) and SAN (130 TB)
- NAS boxes have 16 EIDE disks and 3Ware controllers
  - problems with 3ware controllers
- SAN cluster file system (GPFS) exported via NFS to the WNs
  - high availability through multiple redundant servers
  - load balancing via automounter program map
  - since introduction of above: CPU/Wall clock time nears 1
- planned offering of (x)rootd on file servers

Storage

- tape library IBM 3584 LTO Ultrium
- 8 drives LTO-1, 4 drives LTO-2
- 250 TB native (uncompressed) available
- Tivoli Storage Manager (TSM) for Backup and Archive
- installation of dCache in progress
  - tape backend interfaced to Tivoli Storage Manager
  - installation with 1 head and 3 pool nodes currently tested by CMS & CDF
- other
  - SAM station caches for D0 and CDF
  - JIM (Job information management) station for D0
  - tape connection via scripts (D0)
  - CORBA Naming service (for CDF)

Tasks for LCG-2 / EGEE

- provide resources for Data Challenges (currently Alice & CMS)
- implement LCG-2 on all WNs
- test and provide the LCG-2 SE with dCache/TSM

- distributed ROC together with DESY, FhG and GSI
- collaboration with about 20 German universities
  - support them with LCG-2 / EGEE installation
    (Univ. Wuppertal was the first to join with LCG-2)
  - spread the usage of GermanGrid-CA certificates
  - provide VO and/or RA support
- extend the LCG Global Grid User Support Centre at GridKa for EGEE needs
- test and deploy methods to prevent firewalls in the Grid

No equipment without people. Thanks !

We appreciate the continuous interest and support by the Federal Ministry of Education and Research, BMBF.

Bundesministerium
für Bildung
und Forschung