



DESY Mass Storage First Steps towards global data GRID

A Story about Abstractions and Interfaces



Nov 18, 2002

Patrick Fuhrmann, DESY

First Steps towards global data GRID



Content



Tutorial I : Tape Storage Hardware

Abstraction : The Storage Manager

Pitfall : The Storage Manager

Tutorial II : The Cache

Abstraction : dCache

The Concept

dCache @ DESY , dCache @ FERMI

Tutorial III : The GRID

Abstraction : The GRID



Tutorial I (Nasty Hardware)

This is a ROBOT (Powderhorn) ->

Holds up to 6000 Cartridges

Does up to 450 mounts/dismounts per hour



We are operating 4 of them

This is a TAPE DRIVE (9840) ->

Takes 20 GB and does 10 MB/sec

Needs 8 seconds to become READY



We are operating 25 of them

<- This is another TAPE DRIVE (9940)

Takes 200 GB and does 30 MB/sec

Needs 60 seconds to become READY

We will get them soon



Abstraction Level 1 : The HSM

The HSM handles



Robots

Tapes

Drives

Broken Devices

High Speed Network

You Get

Datasets

&

Perfectly Normal File System

&

Copy Program (osmcp)



Pitfall 1 : The HSM

Even the most sophisticated abstraction doesn't improve the specification of your system

Requirements : 100 TB permanent storage / Year

Current Technology

1 Robot / Year

\$ 1/2 million / Year


new Technology (9940)

\$ 1/2 million / 10 Years



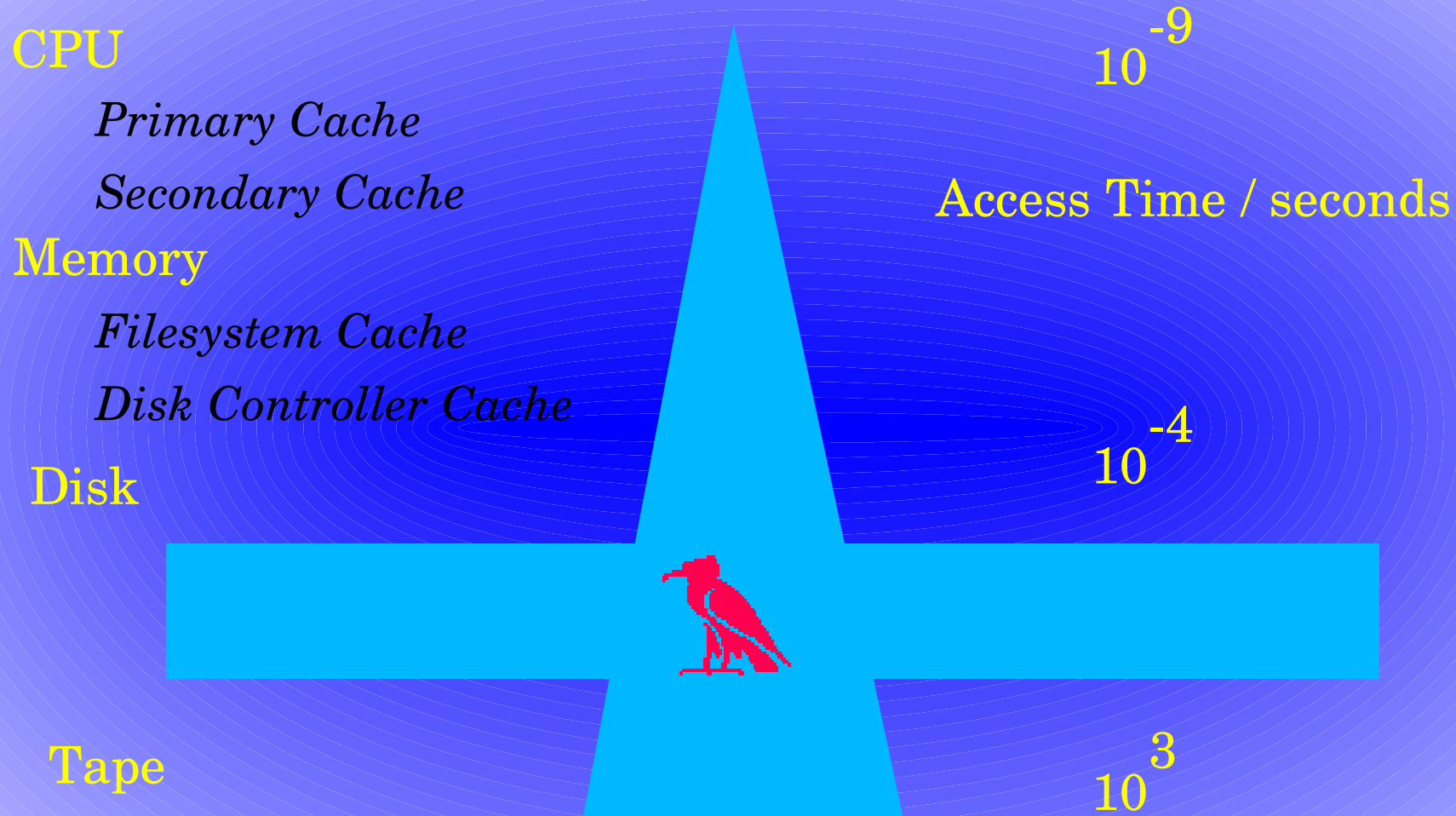
Pitfall 1 : The HSM (cont.)

BUT

- * Tapesystem : bunch of non shareable devices
Robot, Drive, Tape
- * High capacity tapes intensify deficiencies :
 - * *more files are locked*
 - * *increased load/unload times (about 5 times)*
- * one drive one client
 - o(10) high I/O client hosts (SGI)*
 -  *o(1000) weak I/O client hosts (Linux)*



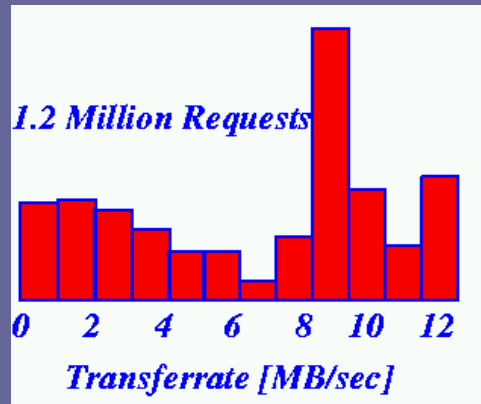
Tutorial II : Caching



Tutorial II : Cache mechanisms

Transfer Speed Adaption

slow client <-> fast drive



Deferred write

*collect data before
dumping to disk*

550,000 Write Requests



20,000 Mounts

Read ahead

*dump whole tape
if one file is requested*

Staging

reuse of cached data



dCache Concept

Standard Mechanisms :

transfer rate adaption

deferred write

staging

read ahead not yet implemented

Interface

Pnfs, dccp

dCache

Pnfs, osmcp

HSM Software (osm, enstore)

Hard & Firmware : Robots , Drives , Tapes



dCache Concept ++

User View

Direct access through shared and preload library

linked to experiment software + Root , Objectivity AMS

Access via mounted Pnfs or dcap://pnfs/desy.de/<exp>/...

Supported Protocols :

dcap & ssl or kerberos (library + copy program)

Kerberized FTP

GRID Ftp

Http

User is never involved in staging the data



dCache Concept ++

System View

Highly distributed server nodes $o(1000)$

Attraction Model

Assigning cache nodes to storage groups (raw99 , dst00 ...)

Support of arbitrarily deep fallback cache nodes

Files can be declared sticky

Cost (cpu , space) dependent request steering

Duplicates on high load to avoid hot spots (pool to pool)

Automatic draining of duplicates on low load



dCache Concept ++

System View

Automatic client library reconnect on any kind of errors

Pools are automatically disabled on disk errors

Manual prestaging (application or API)

Retry or hold on HSM problems (not a client error)

Continuous Disk Sweeper (remove candidates)



dCache & DESY I

Support model

User groups buy tape and disk space from IT

Some Numbers @ DESY

28 TBytes of cache space in total

2 - 10 TBytes are delivered to the clients per day

50,000 - 160,000 files per day

Up to 6 files /second

5 % - 30 % reload from tape



dCache @ FERMI I

Decent Fermi dCache users :

- * MINOS (Neutrino) -> GFtp from Soudan mine.
- * MiniBoone (Neutrino) -> GFtp and dCap
- * Auger (Cosmic Rays) -> GFtp from Argentina and France
- * Grid Condor Project (High Throughput Computing)
NeST -> SRM -> dCache -> Enstore



dCache @ FERMI II

Heavy Fermi dCache users :

- CMS

10 TBytes of Disk Storage

Measured 100 MBytes / sec throughput

- CDF

85 TBytes of (xfs) space within the next two months

Upgrading farmnodes to be able to use the dCache

GRid FTP delivers data to Chicago and Rutgers Uni (new Jersey)

Developing a SAM (sam-cache) interface to dCache

Considering to write raw data through the dCache



Tutorial III : The Grid

Definition I (Kesselman & Foster) 1998 :

A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.

Definition II (Foster, Tuecke) 2000 :

Coordinated resource sharing and problem solving in dynamic, multi institutional virtual organizations.

Definition III (Foster) 2002 :

- 1) Coordinates resources that are not subject of centralized control
- 2) using standard, open, general-purpose protocols and interfaces
- 3) to deliver non trivial qualities of service

Definition IV 2002 :

Money making machine



Tutorial III : The Grid , the goal

What you need to do :

prepare a computing job

specify the datasets you need

specify the money you want to spend / time you want to wait

What the GRID does (in theory)

splits up your job into independent subtasks

finds your datasets (replication cat.)

matchmaker tries to find compromise between data,cpu,network

makes datasets available (offline -> nearline -> online)

starts and monitors your job (not to forget billing)

collects the results and makes them available to you



Tutorial III : The Grid , eu-grid status

Organization OK X (alarms)

Mon, 18 Nov 2002 11:39:12 GMT (refresh=10min)

Map Link Symbol No Status Normal TCP/UDP/URL failed Ping failed

- CERN** (alarms)
 - Production** (alarms)
 - Development** (alarms) up<=
- PPARC** (alarms)
 - RAL** (alarms)
 - Production** (alarms)
 - Developpement** (alarms) up<=
 - Manchester** (alarms)
 - Bristol** (alarms)
 - Glasgow** (alarms)
 - Birmingham** (alarms)
 - QMW** (alarms)
 - ICSTM** (alarms)
 - RHUL** (alarms)
- INFN** (alarms)
 - CNAF** (alarms)
 - Production** (alarms)
 - Development** (alarms) up<=
 - Cagliari** (alarms)
 - Catania** (alarms)
 - Milano** (alarms)
 - Padova** (alarms)
 - Pisa** (alarms)
 - Torino** (alarms)
 - Legnaro** (alarms)
 - Roma 1** (alarms)
 - Roma 2** (alarms)

68 Monitored Farms



Tutorial III : The Grid , problems

Reality :

Splitting up into subjobs doesn't work yet

Resource Broker Info becomes out of date

Certificates have to be matched to virtual organisations

Europe doesn't except globus.org certificates

US-DEO doesn't except european certificates

Only supported platform is a certain redhat version

Weak or missing interface definition to the local fabrics



Tutorial III : The Grid

GriPhyN

Grid Physics Network



iVDGL

Int. Virtual Data Grid Lab



Globus



DataTAG



Nov 18, 2002

Patrick Fuhrmann, DESY

First Steps towards global data GRID



Tutorial III : The Grid

Needed during a year of LHC operations

Tape	Disk	CPU
29'400 TB	9'600 TB	$6.2 * 10^6$ SI95

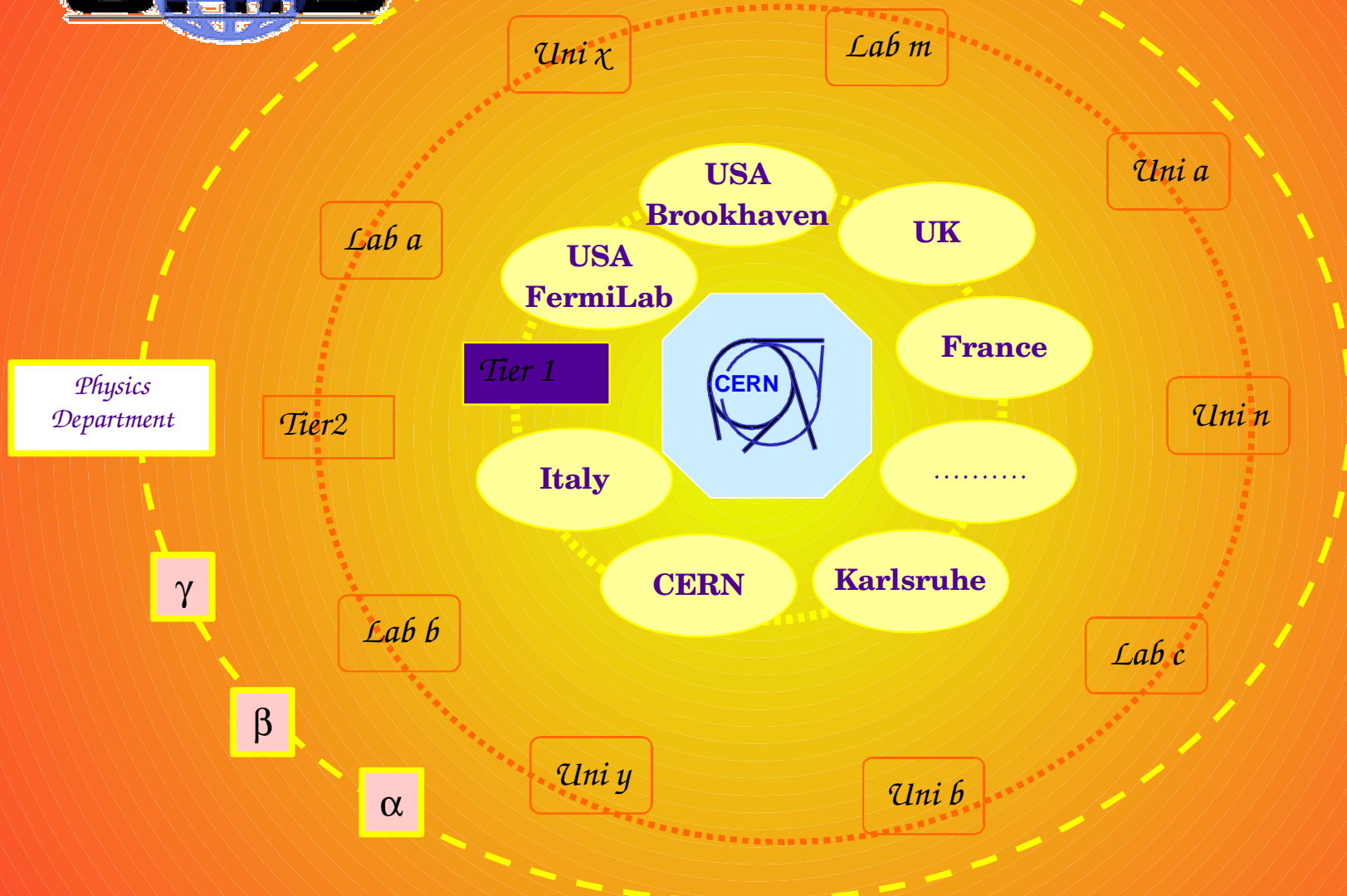
In today's units:

60 STK Silos	160'000 60GB disks	150'000 800 MHz CPUs
-----------------	-----------------------	-------------------------

Taken from: LHC Computing Review, CERN/LHCC/2001-004



Tutorial III : The Grid



(stolen from Ingo Augustin CERN , Karlsruhe Oct 30, 2002)



Nov 18, 2002

Patrick Fuhrmann, DESY

First Steps towards global data GRID

Tutorial III : The Grid LHC World



LHC: > 5000 physicists
> 270 institutes
> 60 countries

(stolen from Ingo Augustin CERN , Karlsruhe Oct 30, 2002)



Nov 18, 2002

Patrick Fuhrmann, DESY

First Steps towards global data GRID

Tutorial III : The Grid Future

Goal : Artificially Producing a Hype (same as web)

Globus Toolkit

Simplifies Interoperability

Global Grid Form (GGF)

Defines Reliable Interfaces

Indispensable for finding partners in Industry

Open Grid Services Architecture (OGSA)



The Mass Storage Fabric Interface

Storage Resource Manager

Missing Link in globus mass storage design

Interface proposed by Fermi, JLab, Argonne

Based on XML (soap) remote procedure calls

Supports queries like

Time to get a file
reserve space

make file online
pin file to disk

transfer protocol negotiation dCap, Gftp, http

SRM implemented in CASTOR Jasmin and dCache

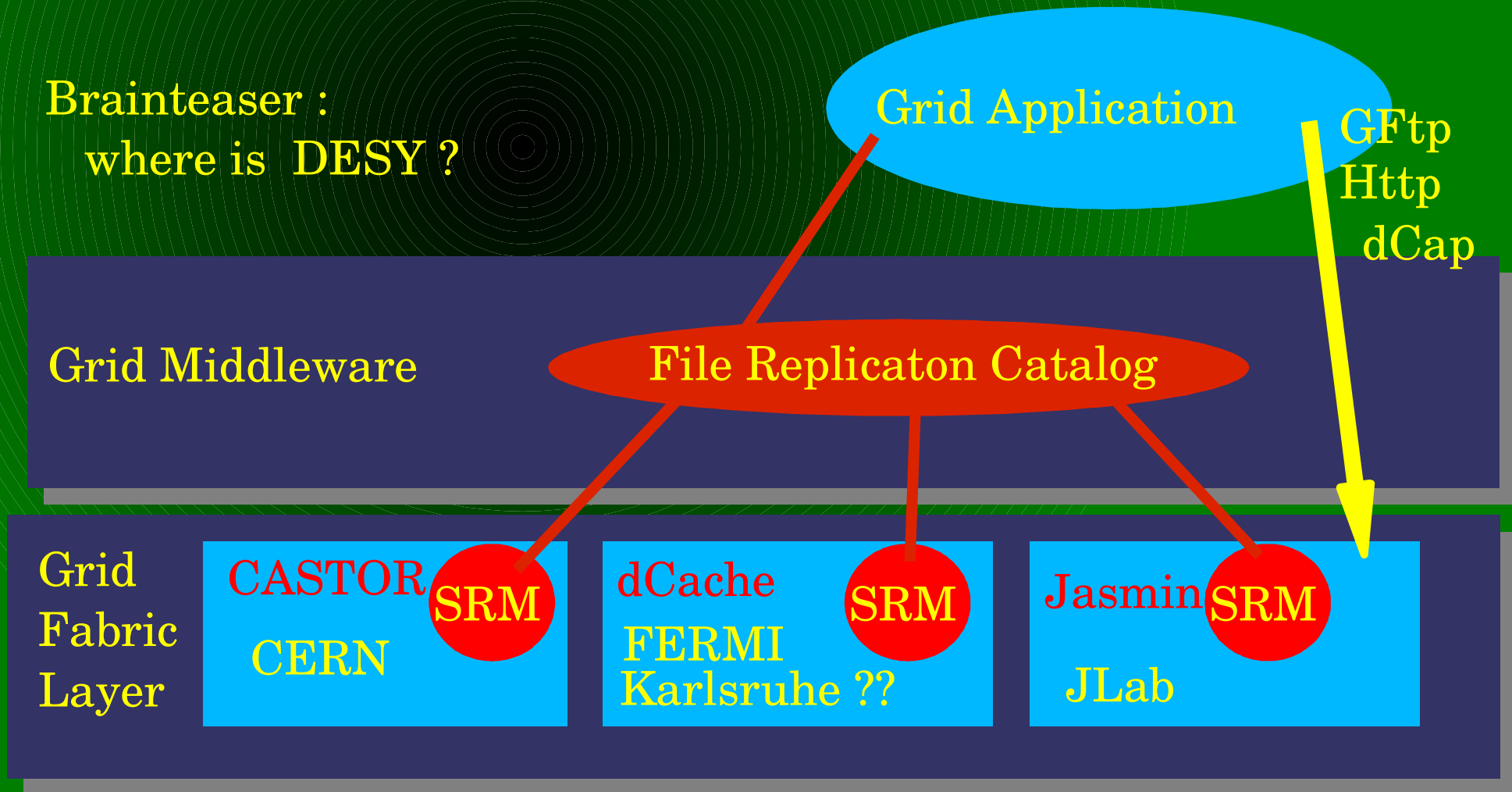
Globus project for the first time presented a slide mentioning an SRM without calling it an SRM.

So still negotiations needed.



Second mass storage abstraction

Brinteaser :
where is DESY ?



Hier ist Ende



Nov 18, 2002

Patrick Fuhrmann, DESY

First Steps towards global data GRID