Providing GRID Data Services TODAY

or: connect an existing data service fabric to a GRID

Patrick Fuhrmann, Martin Gasthuber **DESY/Hamburg**
Rich Wellner **Fermi**

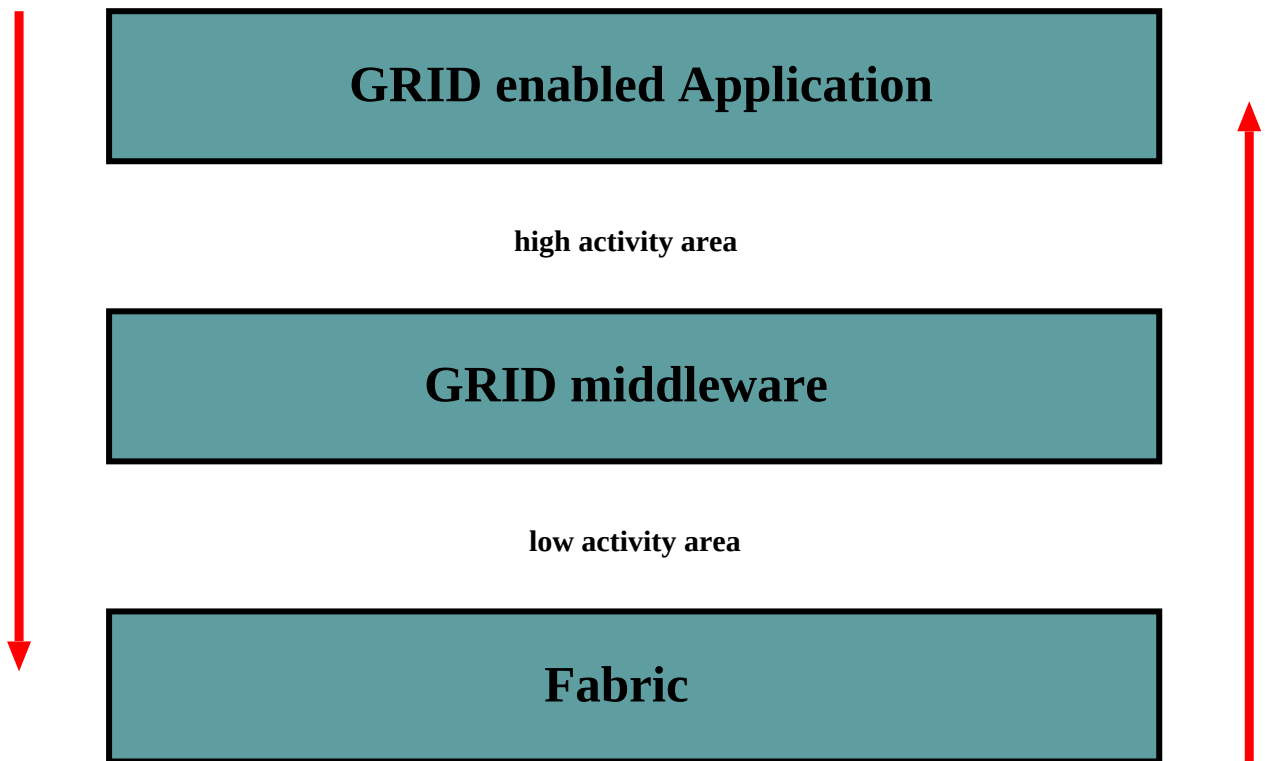
Initial Remarks

- We only talk about the data management, data access aspect of GRID computing
  - the 'Fabric' here is an existing data management (services) fabric
    running @ DESY and Fermi

- We are not in deep touch to any of the current (HEP) GRID projects
  - Fermi has (active) contacts to Globus @ Argonne

- Goal
  connect existing (data) fabric to applications (or vice versa)
  through a GRID middleware

the MAP


GRID enabled Application
high activity area
GRID middleware
low activity area
Fabric

The existing data services - what is it

- DESY and Fermi use the same architecture and mostly the same components
  - for the nameservice: **PNFS**
  - for the tape service: **Enstore (Fermi), OSM (Desy)**
  - for the disk caching: **dCache (collaboration DESY/Fermi)**

- Providing
  - single uniform (and scalable) namespace for all files in the system
    - independent of file location and multiplicity
    - use NFS v2 protocol to access
    - efficient, scalable internal architecture
  - simple, scalable tape services
    - direct access possible
  - distributed disk cache
    - dramatic reduction of tape load
    - direct - random access from application - using shared lib preload
    - flexible management / data flow steering
    - policy based data placement

scale in admin

Perfectly Normal FS

* supports 8 layers per file entry.
  - Top layer appears as regular File Entry
    but denies I/O
  - Others can are hidden
    but can be accessed by ’spooky’ filenames
* supports inheritable directory tags.
* supports wormholes.
* uses eventhandler for ’remove’ and ’move’.
* is unaware of HSM or dCache.

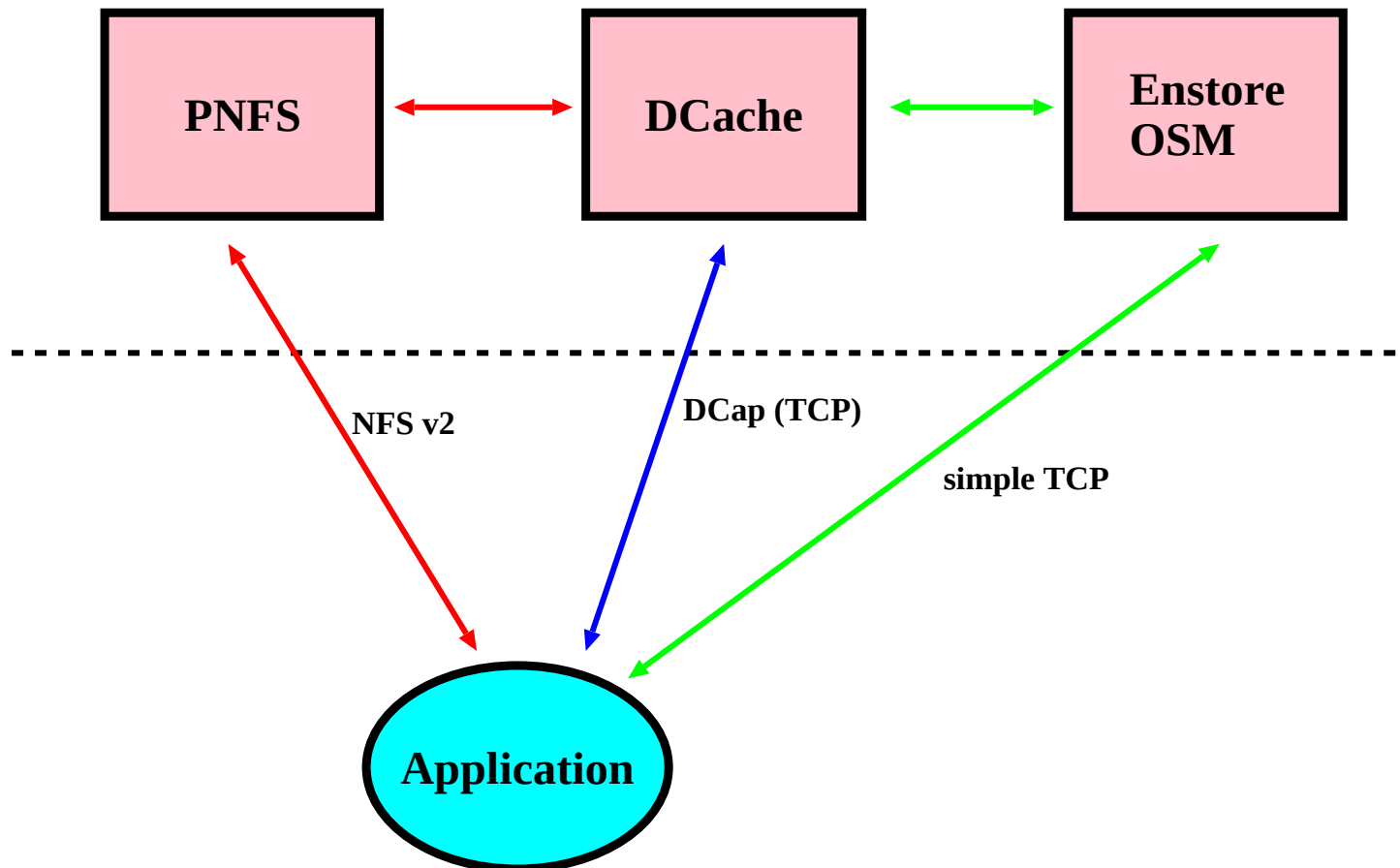
dCache (1)

- Failsafe - reconnect, re-stage - hidden to application
- Data placement policy expressed by:
  - network topology (source, destination) include masks
  - storage group (set of physical tape media)
  - PNFS directory tag
  - Costs (CPU Load & (free) Space)
- Sticky bit - pin files on cache pools
- Thread safe access library (.a and .so)
- Rules for pool selection - always fallbacks available
- Secure/delegated management and administration
  - ACLs
  - Kerberos based authentication
  - ssh secured login (i.e. just use the ssh application)
  - WEB monitoring
- ROOT interface available (initial version)
  - room for performance enhancements

dCache (2)

- Pool to Pool transfers
  - immediate read after write (file not on tape yet)
  - file replication on different pools
- Authenticated/secure data access (control line)
  - Kerberos 5 (GSSAPI)
  - SSL
- Configurable load limits
  - # of movers (active transfers to/from client)
  - # of HSM stores/restores
- Pluggable Protocol Engines
  - HSM -> Enstore, OSM, ...
  - Client -> dCap, FTP, ... (GRID access protocols)
- URL based access - no NFS mount required
  - dc_open("dcap://dcachedoor.desy.de/pnfs.desy.de/zeus/...", O_READ)
- 100% made of Java (same code on Linux,IRIX,Solaris)
- 100% free of experiment code (real generic)

Simple Picture



dCache Components

Access Points

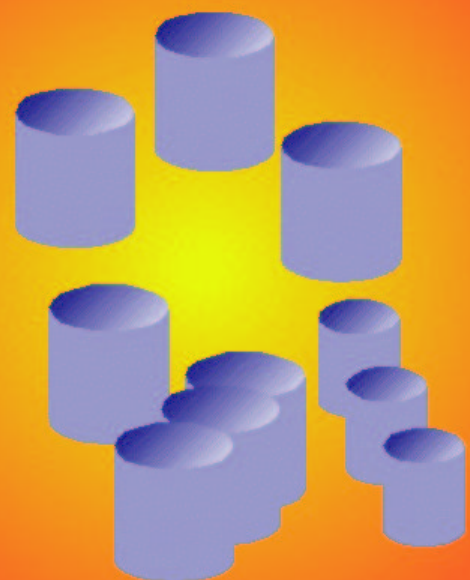
Grid

- Grid FTP
- Grid Fabric Interface

Native Access

- dCap

dCache Kernel



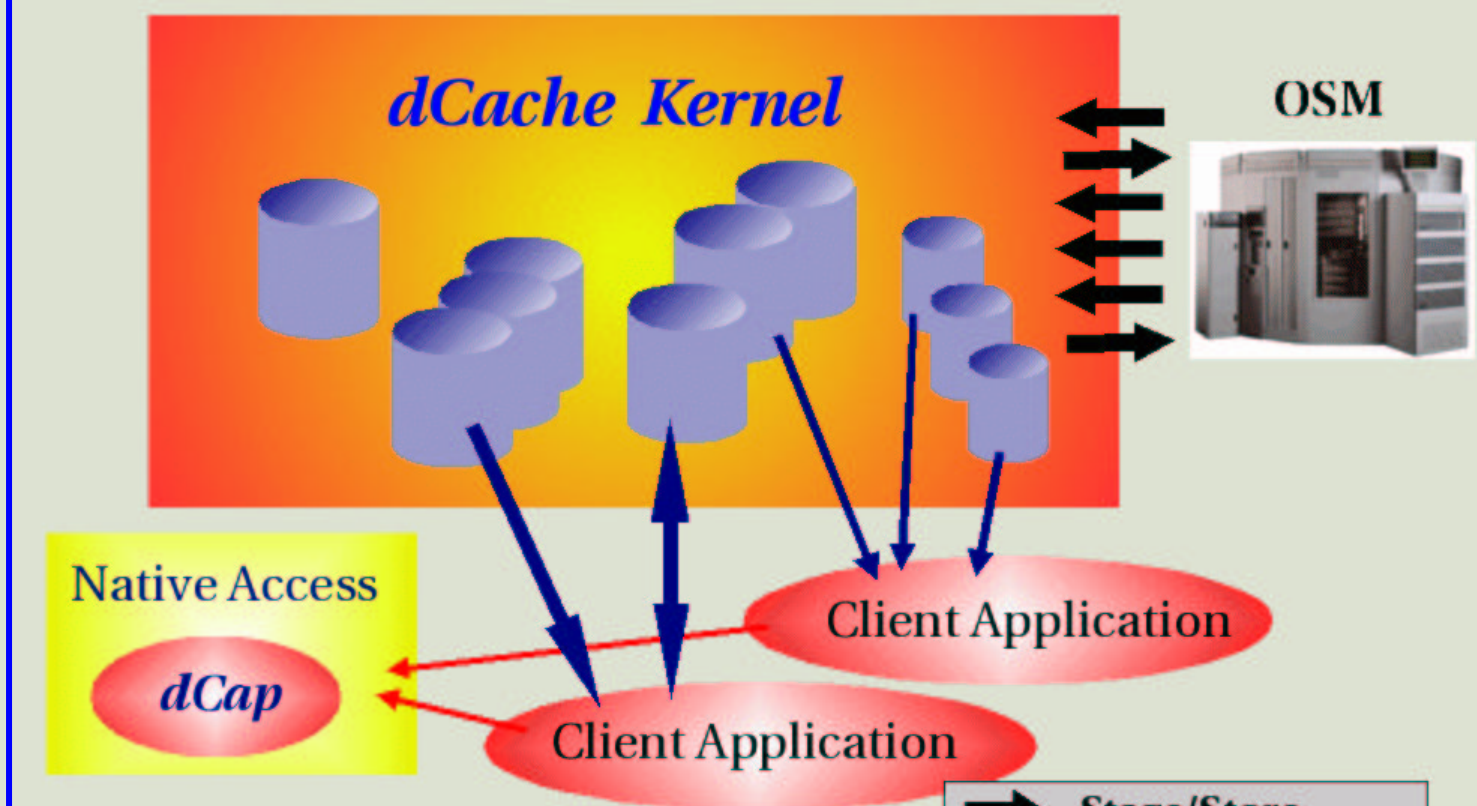
Storage Manager

OSM

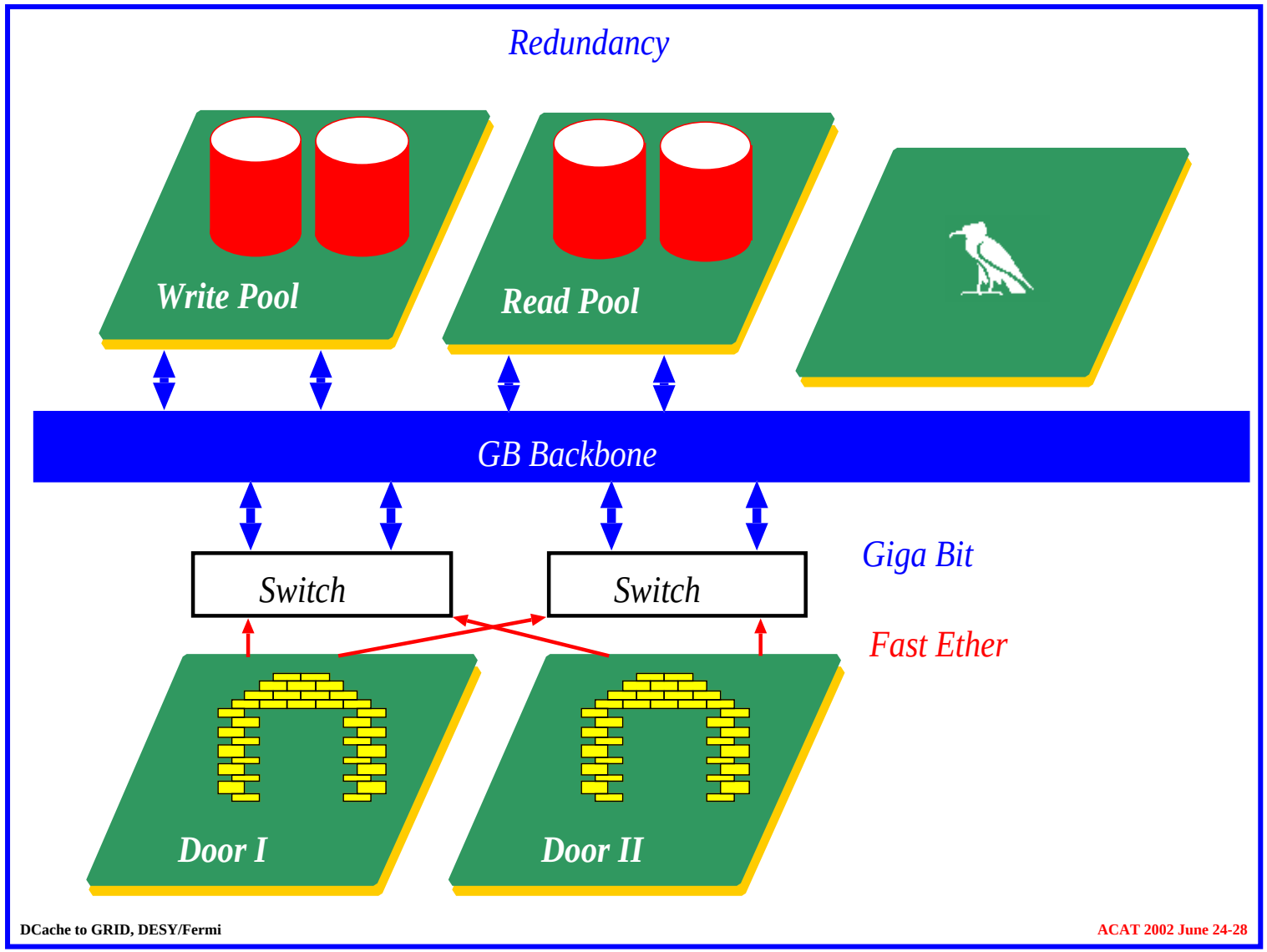
Enstore

Patrick Fuhrmann, Desy IT, dCache Project

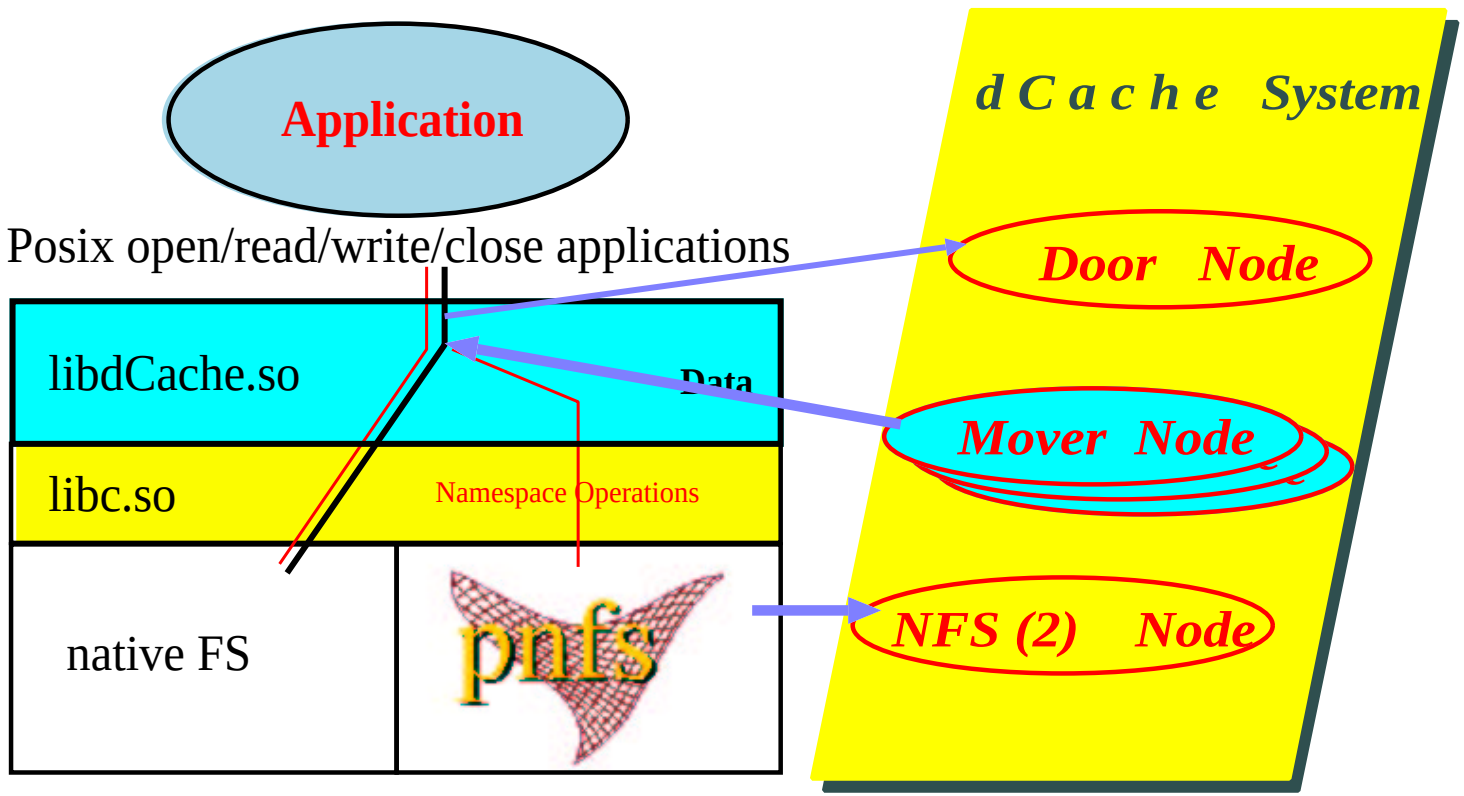
dCache Native Access

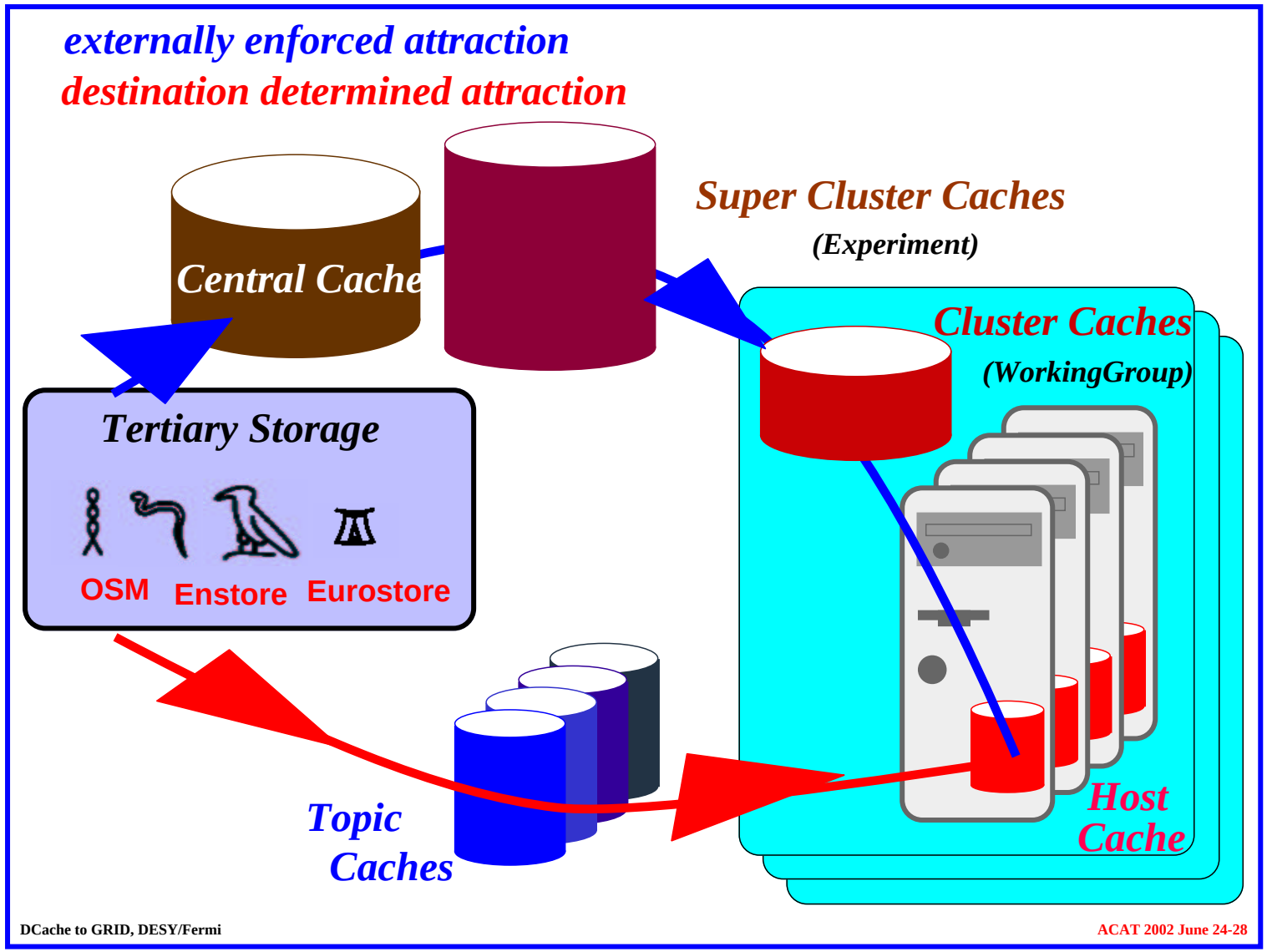


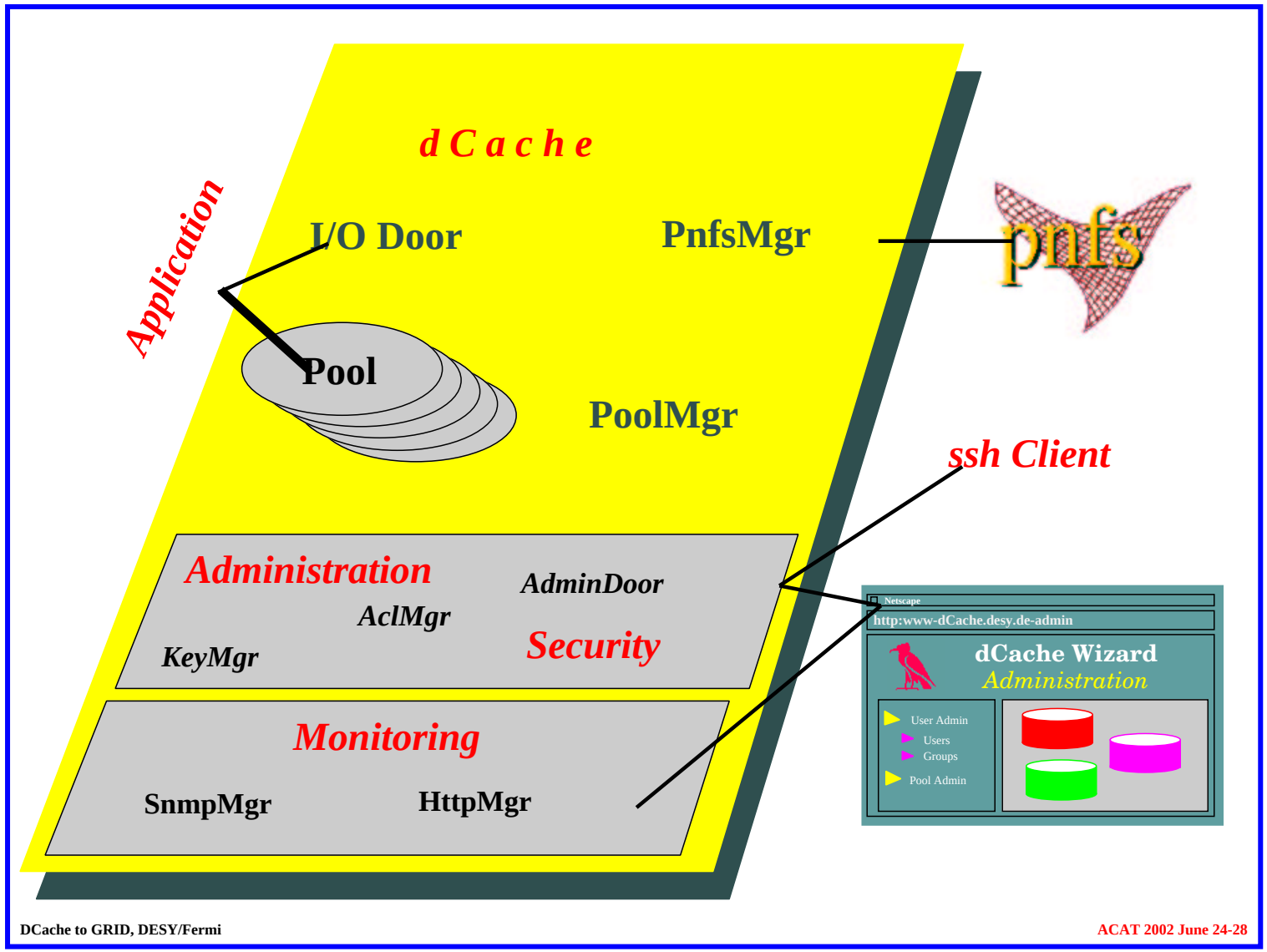
Patrick Fuhrmann, Desy IT, dCache Project


Redundancy
Write Pool
Read Pool
GB Backbone
Switch
Switch
Giga Bit
Fast Ether
Door I
Door II

The dCap Library




externally enforced attraction
destination determined attraction
Central Cache
Super Cluster Caches
(Experiment)
Cluster Caches
(WorkingGroup)
Tertiary Storage
OSM
Enstore
Eurostore
Topic
Caches
Host
Cache


dCache
Application
I/O Door
PnfsMgr
pnfs
Pool
PoolMgr
ssh Client
Administration
AdminDoor
AclMgr
KeyMgr
Security
Monitoring
SnmpMgr
HttpMgr
Netscape
http:www-dCache.desy.de-admin
dCache Wizard
Administration
User Admin
Users
Groups
Pool Admin

dCache Numbers for June 2002

Total Repository (June 20) : 15.0 Tbytes

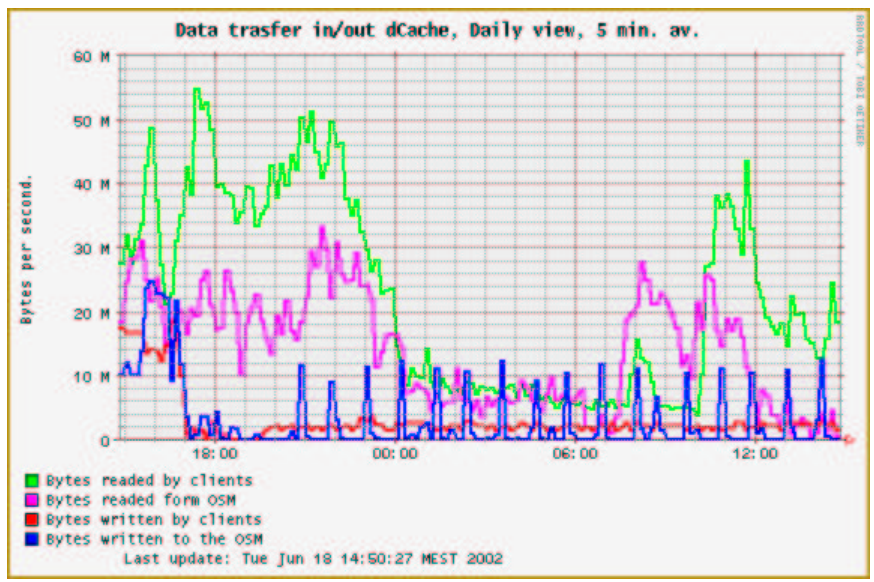
dCache --- >>>> OSM : 0.5 TBytes / day

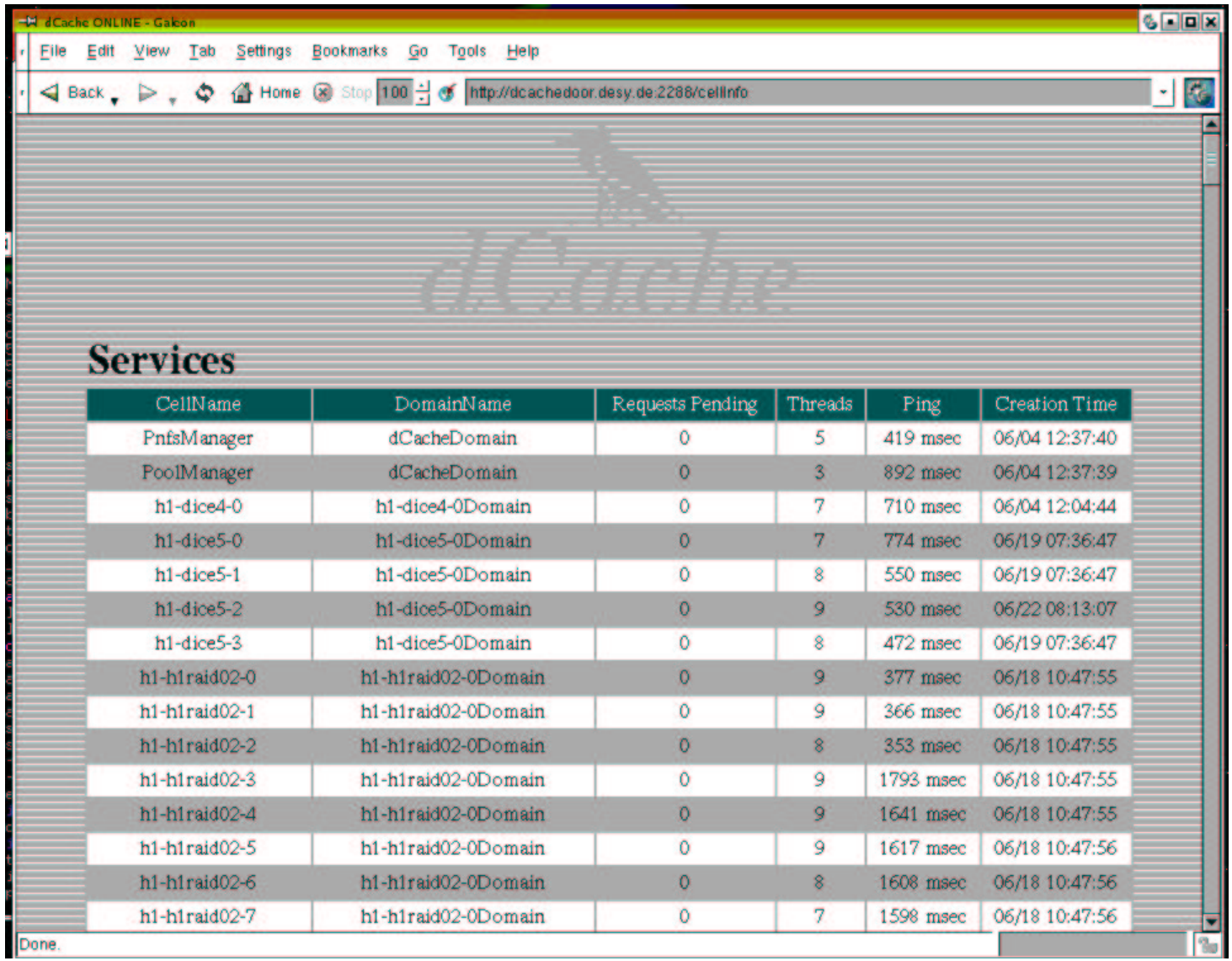
OSM --->>>> dCache : 1.0 TBytes / day

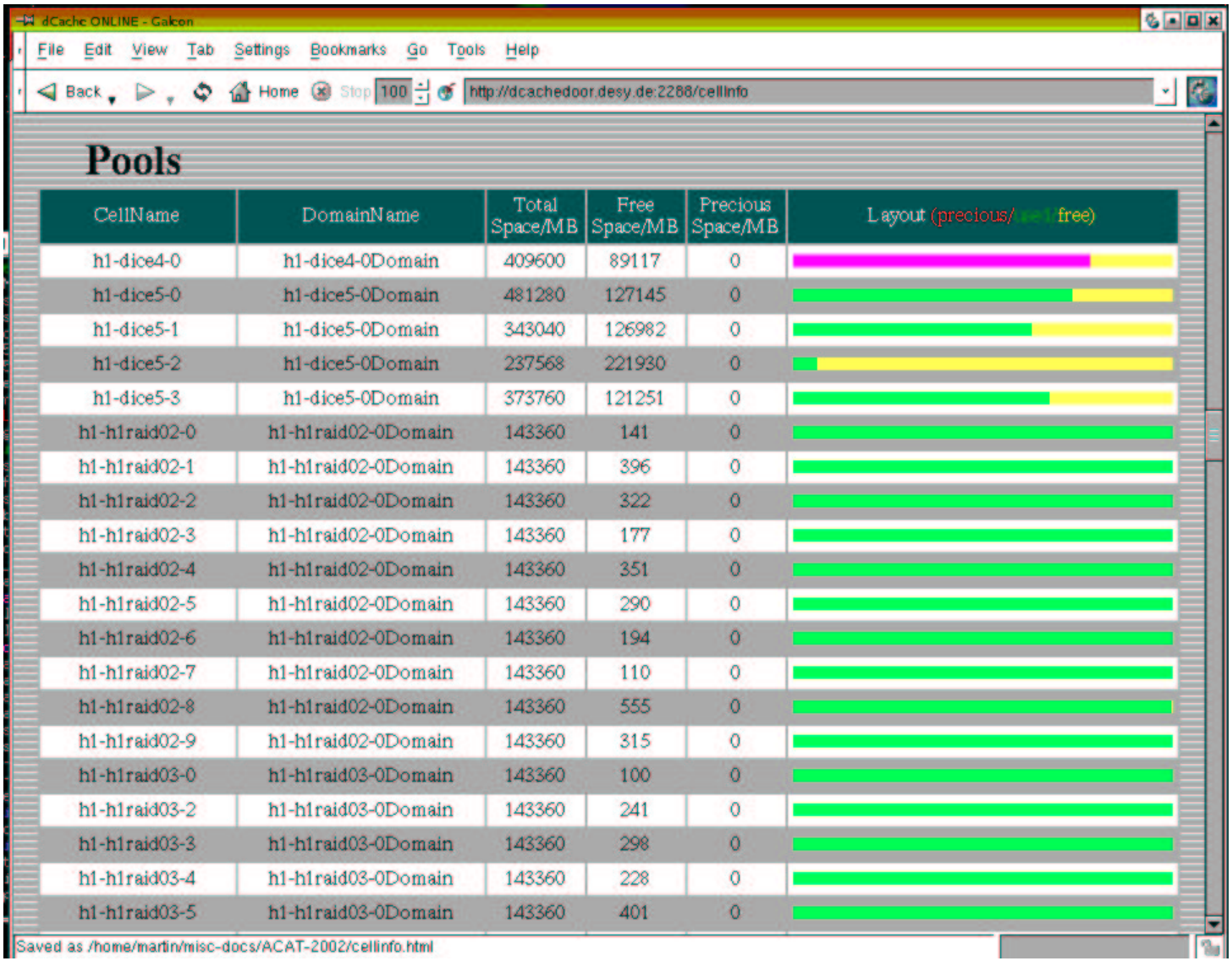
dCache <<<->>> clients : 2.7 TBytes / day

Maximum observed request rate : 8 / second

Patrick Fuhrmann, Desy IT, dCache Project


Data trasfer in/out dCache, Daily view, 5 min. av.
Bytes per second.
60 M
50 M
40 M
30 M
20 M
10 M
0
18:00
00:00
06:00
12:00
Bytes readed by clients
Bytes readed form OSM
Bytes written by clients
Bytes written to the OSM
Last update: Tue Jun 18 14:50:27 MEST 2002


dCache ONLINE - Galeon
File Edit View Tab Settings Bookmarks Go Tools Help
Back Home Stop 100 http://dcachedoor.desy.de:2288/cellInfo
dCache
Services
CellName DomainName Requests Pending Threads Ping Creation Time
PnfsManager dCacheDomain 0 5 419 msec 06/04 12:37:40
PoolManager dCacheDomain 0 3 892 msec 06/04 12:37:39
h1-dice4-0 h1-dice4-0Domain 0 7 710 msec 06/04 12:04:44
h1-dice5-0 h1-dice5-0Domain 0 7 774 msec 06/19 07:36:47
h1-dice5-1 h1-dice5-0Domain 0 8 550 msec 06/19 07:36:47
h1-dice5-2 h1-dice5-0Domain 0 9 530 msec 06/22 08:13:07
h1-dice5-3 h1-dice5-0Domain 0 8 472 msec 06/19 07:36:47
h1-h1raid02-0 h1-h1raid02-0Domain 0 9 377 msec 06/18 10:47:55
h1-h1raid02-1 h1-h1raid02-0Domain 0 9 366 msec 06/18 10:47:55
h1-h1raid02-2 h1-h1raid02-0Domain 0 8 353 msec 06/18 10:47:55
h1-h1raid02-3 h1-h1raid02-0Domain 0 9 1793 msec 06/18 10:47:55
h1-h1raid02-4 h1-h1raid02-0Domain 0 9 1641 msec 06/18 10:47:55
h1-h1raid02-5 h1-h1raid02-0Domain 0 9 1617 msec 06/18 10:47:56
h1-h1raid02-6 h1-h1raid02-0Domain 0 8 1608 msec 06/18 10:47:56
h1-h1raid02-7 h1-h1raid02-0Domain 0 7 1598 msec 06/18 10:47:56
Done.


dCache ONLINE - Galeon
File Edit View Tab Settings Bookmarks Go Tools Help
Back Home Stop 100
http://dcachedoor.desy.de:2286/cellInfo
Pools
CellName | DomainName | Total Space/MB | Free Space/MB | Precious Space/MB | Layout (precious/free)
h1-dice4-0 | h1-dice4-0Domain | 409600 | 89117 | 0
h1-dice5-0 | h1-dice5-0Domain | 481280 | 127145 | 0
h1-dice5-1 | h1-dice5-0Domain | 343040 | 126982 | 0
h1-dice5-2 | h1-dice5-0Domain | 237568 | 221930 | 0
h1-dice5-3 | h1-dice5-0Domain | 373760 | 121251 | 0
h1-h1raid02-0 | h1-h1raid02-0Domain | 143360 | 141 | 0
h1-h1raid02-1 | h1-h1raid02-0Domain | 143360 | 396 | 0
h1-h1raid02-2 | h1-h1raid02-0Domain | 143360 | 322 | 0
h1-h1raid02-3 | h1-h1raid02-0Domain | 143360 | 177 | 0
h1-h1raid02-4 | h1-h1raid02-0Domain | 143360 | 351 | 0
h1-h1raid02-5 | h1-h1raid02-0Domain | 143360 | 290 | 0
h1-h1raid02-6 | h1-h1raid02-0Domain | 143360 | 194 | 0
h1-h1raid02-7 | h1-h1raid02-0Domain | 143360 | 110 | 0
h1-h1raid02-8 | h1-h1raid02-0Domain | 143360 | 555 | 0
h1-h1raid02-9 | h1-h1raid02-0Domain | 143360 | 315 | 0
h1-h1raid03-0 | h1-h1raid03-0Domain | 143360 | 100 | 0
h1-h1raid03-2 | h1-h1raid03-0Domain | 143360 | 241 | 0
h1-h1raid03-3 | h1-h1raid03-0Domain | 143360 | 298 | 0
h1-h1raid03-4 | h1-h1raid03-0Domain | 143360 | 228 | 0
h1-h1raid03-5 | h1-h1raid03-0Domain | 143360 | 401 | 0
Saved as /home/martin/misc-docs/ACAT-2002/cellinfo.html

And now - the GRID comes along

- Fermi require data exchange with Institutes in the UK
  - already have Kerberized FTP access

- What interface to implement ??
  - beside GridFTP nothing else common (as we know)

- Within our community we know about:
  1. GridFTP (invented by Globus)
  2. SRM (invented by LBL, JLab, Fermi)

- GridFTP (with parallel streams) implemented
  - includes Kerberos authentification
  - GSI (certificate based) in work

SRM - Storage Resource Manager

- SRM v1 implementation in progress
  - Pre-Allocation (semi persistent) requires more work
  - hope for less internal changes in dCache
  - Demo for next GGF scheduled

  - JLab has done already for JASMine
  - LBL has done already for HPSS
- SRM v1 noticed by Globus
- SRM v2 definition nearly completed
  - by LBL, JLab, Fermi, CERN-EDG
  - We still have a few concerns - lets see
  - see http://sdm.lbl.gov/srm/documents/joint.docs/ SRM.v2.0.joint.func.design.doc

GridFTP

- GridFTP (with parallel streams) implemented
  - includes Kerberos authentification
  - GSI (certificate based) in work

- major problem with FTP port management
  - need changes for clean/scalable implementation
    - GridFTP v2 in preparation (too long)
    - maybe intermitten v1.5 definition
    - joint effort with Globus required

- Demo at next GGF scheduled

Conclusions

- need real running experience (SRM)
  - to get confident in SRM completeness
  - optimize implementation (on both sides)

- GridFTP needs some ’adjustments’

- More (concurrent) interfaces might help
  - before its too late

http://www-dcache.desy.de
http://www-pnfs.desy.de
http://http://www-hppc.fnal.gov/enstore