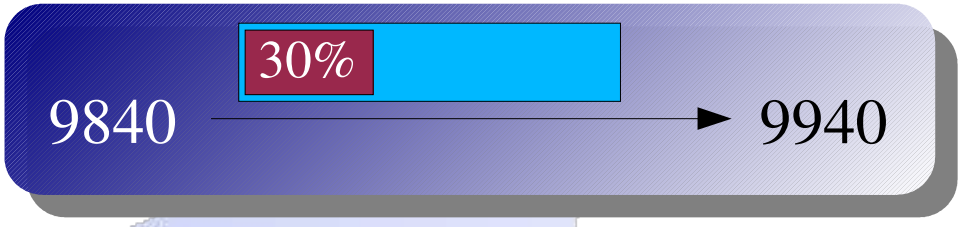
dCache

The Big Picture

Patrick Fuhrmann et al.

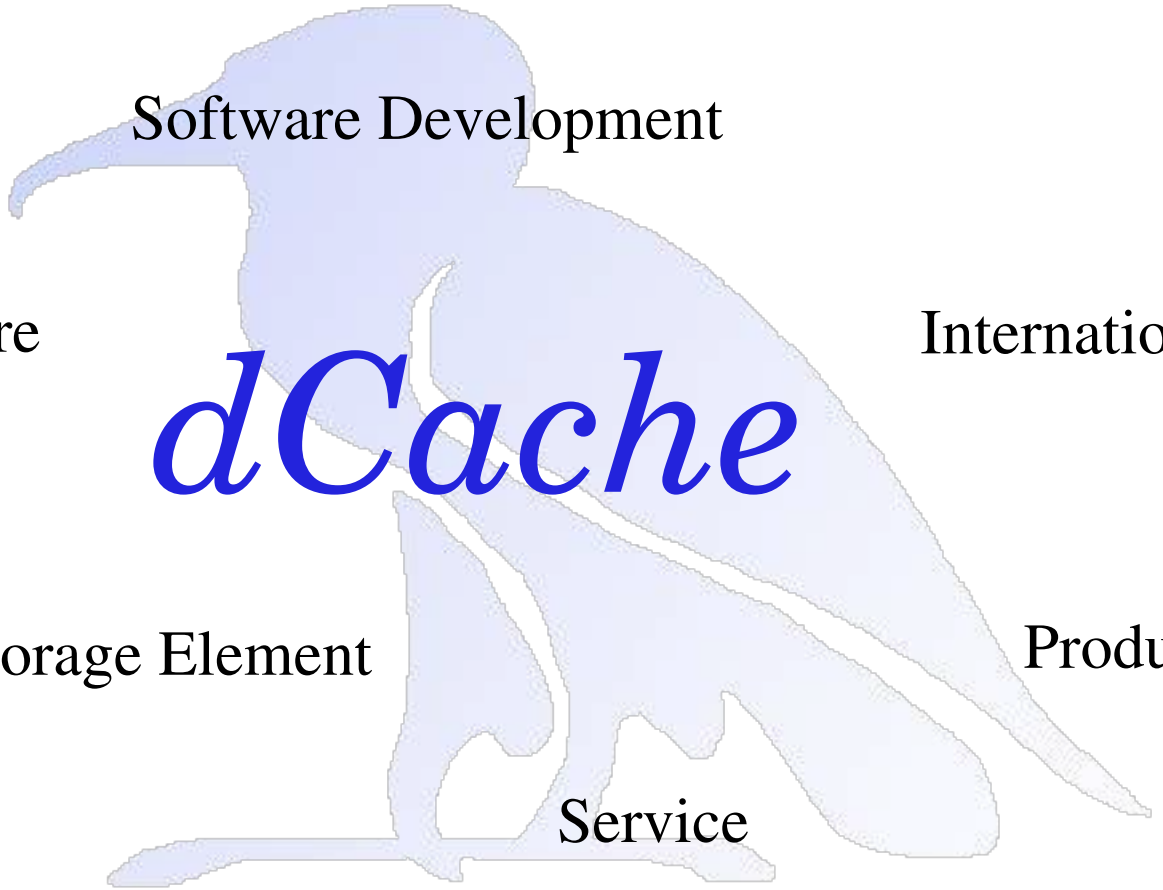
Tape Storage Media Migration @ DESY-HH



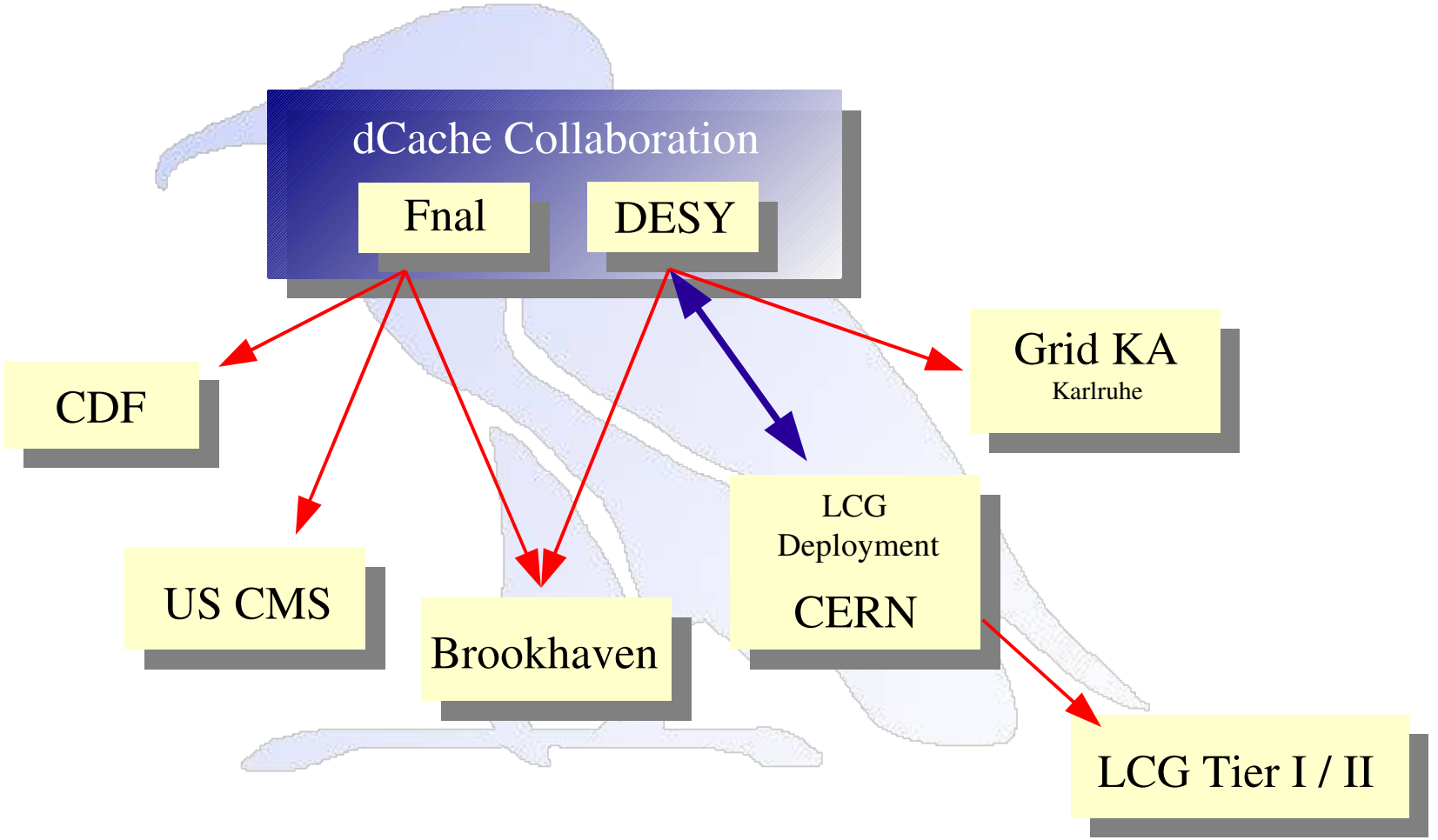
	9840	9940	'titanuim'
Capacity	20 GB	200 GB	500 GB
Transfer Speed (v)	10 MB/sec	30 MB/sec	120 MB/sec
'Ready' Time (T)	6 sec	60 sec	> 60 sec
T * v	60	1800	7800

$$\text{efficiency} = \frac{1}{1 + \frac{T * v}{n * Fs}}$$

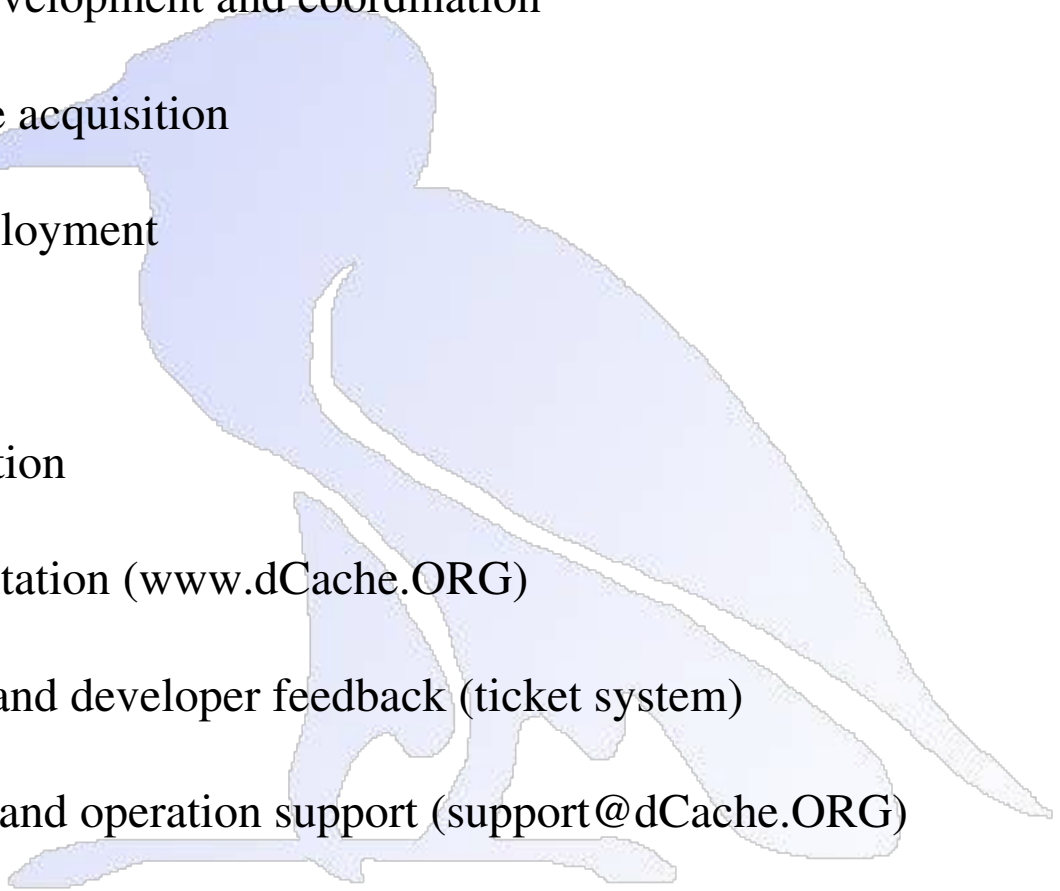
Fs : average file size
n : average files / mount


Software Development
Future
International Project
dCache
Grid Storage Element
Product
Service

dCache is a joined effort between the Deutsches Elektronen Synchrotron (DESY) and the Fermi National Laboratory (FNAL)



Project, Product



- Software development and coordination
- New feature acquisition
- Product deployment
- Packaging
- Documentation
- Web Presentation (www.dCache.ORG)
- Bug report and developer feedback (ticket system)
- Installation and operation support (support@dCache.ORG)

Core

- Combines several hundred pool nodes and lets them look like a single huge file system space.
- Support multiple internal and external copies of a singe file system entry point.
- Performs automatic pool to pool copies of datasets to flatten data access hot spots.
- Fine grained pool selection (experiment, read-write, internal external, priority)
- Cached data only removed if space is running short (no threshold)
- Powerful administration interface via 'ssh' and GUI.
- Scales due to multiple doors.

Resilient Module

- Takes care that at least 'n' but not more than 'm' copies of a single dataset exists within one dCache instance.
- Takes care that this rule is still true if nodes go down (schedules or even unexpected)

HSM connection module (tape access optimization)

- Groups incoming datasets according to HSM specific sorting criteria and flushes them to one or more Tape systems, following certain rules.
- Removes 'old' files from disk, but only if space is running short.
- Retrieves dataset from tape to disk if dataset is requested by dCap/Ftp/Srm open operation without user/administrator interaction.

Supported Access Methods

- Local name space operations via nfs 2/3
- dCap protocol for local area posix like access (plain,kerberos,ssl,gsi)
- Ftp protocol (plain,gsi)
- Storage Resource Manager (SRM)

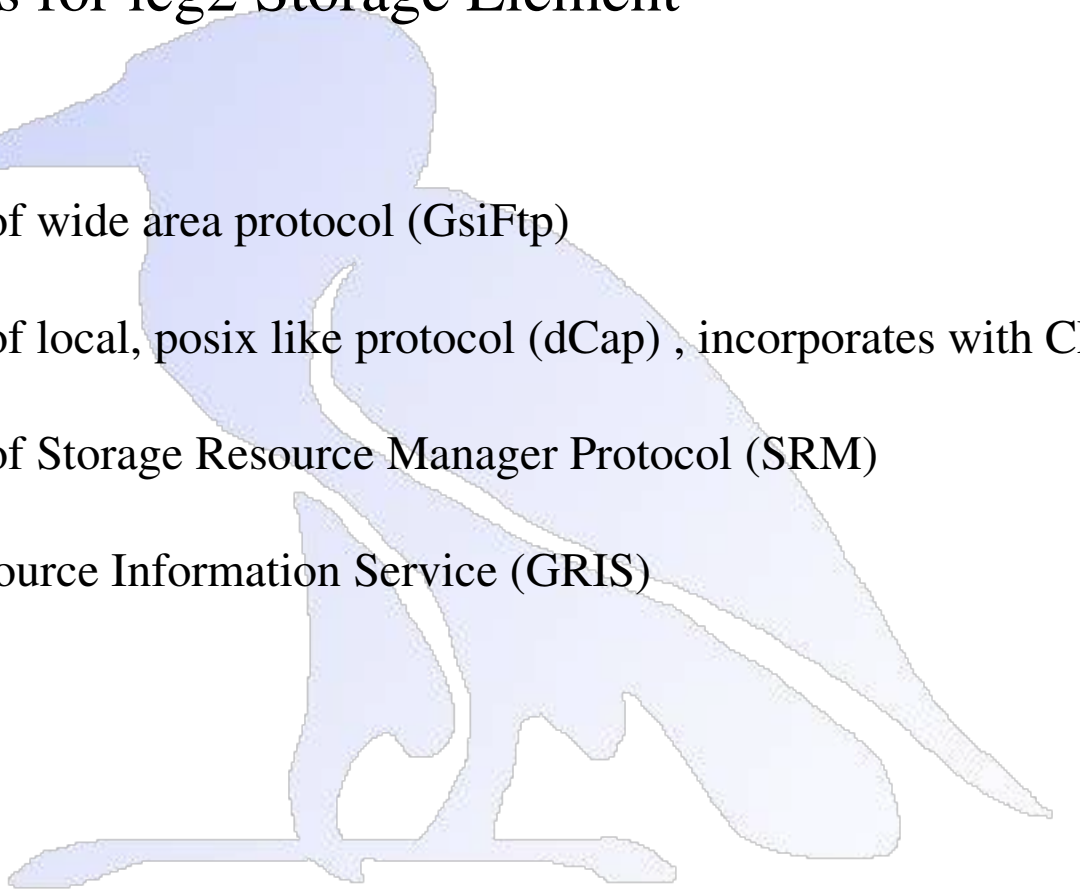
dCap protocol/implementation details

- Supports optimized I/O and name space operations via URL like syntax
- c-language library implementation including PRELOAD
- ls -l dcap://pnfs/desy.de/it/users/patrick
- Supports linux (32 + 64 bit), solaris, (limited windows)
- automatic reconnect on pool or server failures
- dCache interfaced to ROOT
- dCache and non dCache I/O transparently handled by dCap library
- dCap interfaced by GFAL (Grid File access library)
- Read Ahead buffering and deferred write
- Supports Gss(Kerberos), Gsi (Grid) and ssl as secure protocols.
- Thread safe

SRM (Storage Resource Manager) details

- Prepares data transfers, checks certificates and permissions.
- Negotiates transfer protocols (dCap,rfio,ftp,http)
- Retries until transfer succeeds
- Space reservation
- Future : quotas

Requirements for lcg2 Storage Element



- Support of wide area protocol (GsiFtp)
- Support of local, posix like protocol (dCap) , incorporates with CERN GFAL.
- Support of Storage Resource Manager Protocol (SRM)
- Grid Resource Information Service (GRIS)

Desy+H1 Installation

World

h1 - farm

h1 resilient dCache

Pool

R

d 2 cache

Desy HH Main cache

70 TBytes

SRM-dCache

SRM

Pools

GsiFtpDoor

Pools

GsiFtpDoor

4 Nodes = 5 TB

High Performance

H1 scratch

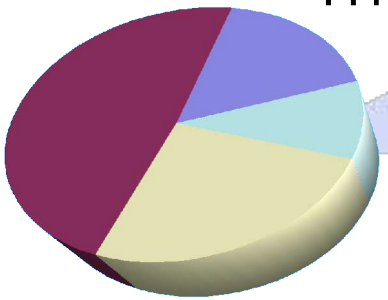
/acsdisk

Desy Main PNFS /acs

Some H1 Statistics (Jan 2005)

dCache disk storage

H1

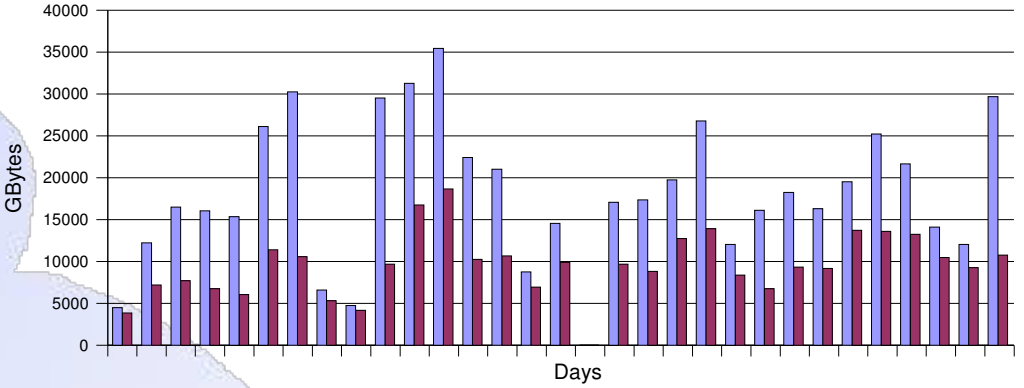


13 Tbytes disk space
211 Tbytes tape space

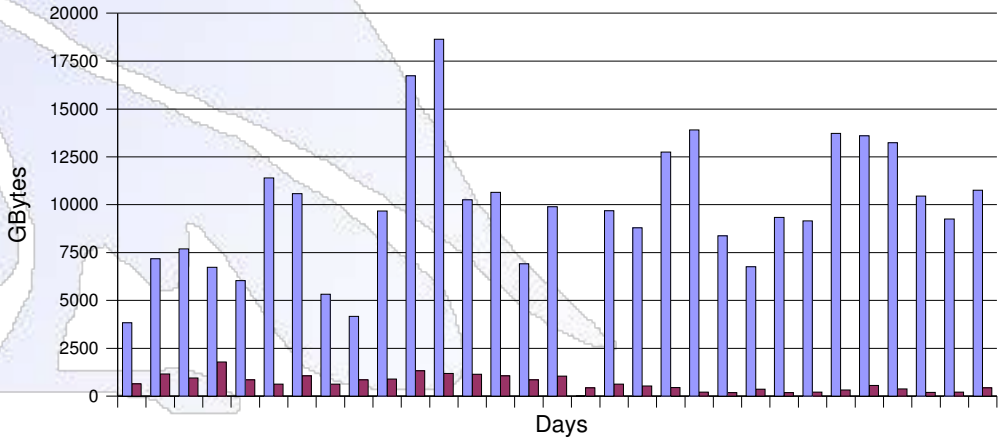
Some other experiment

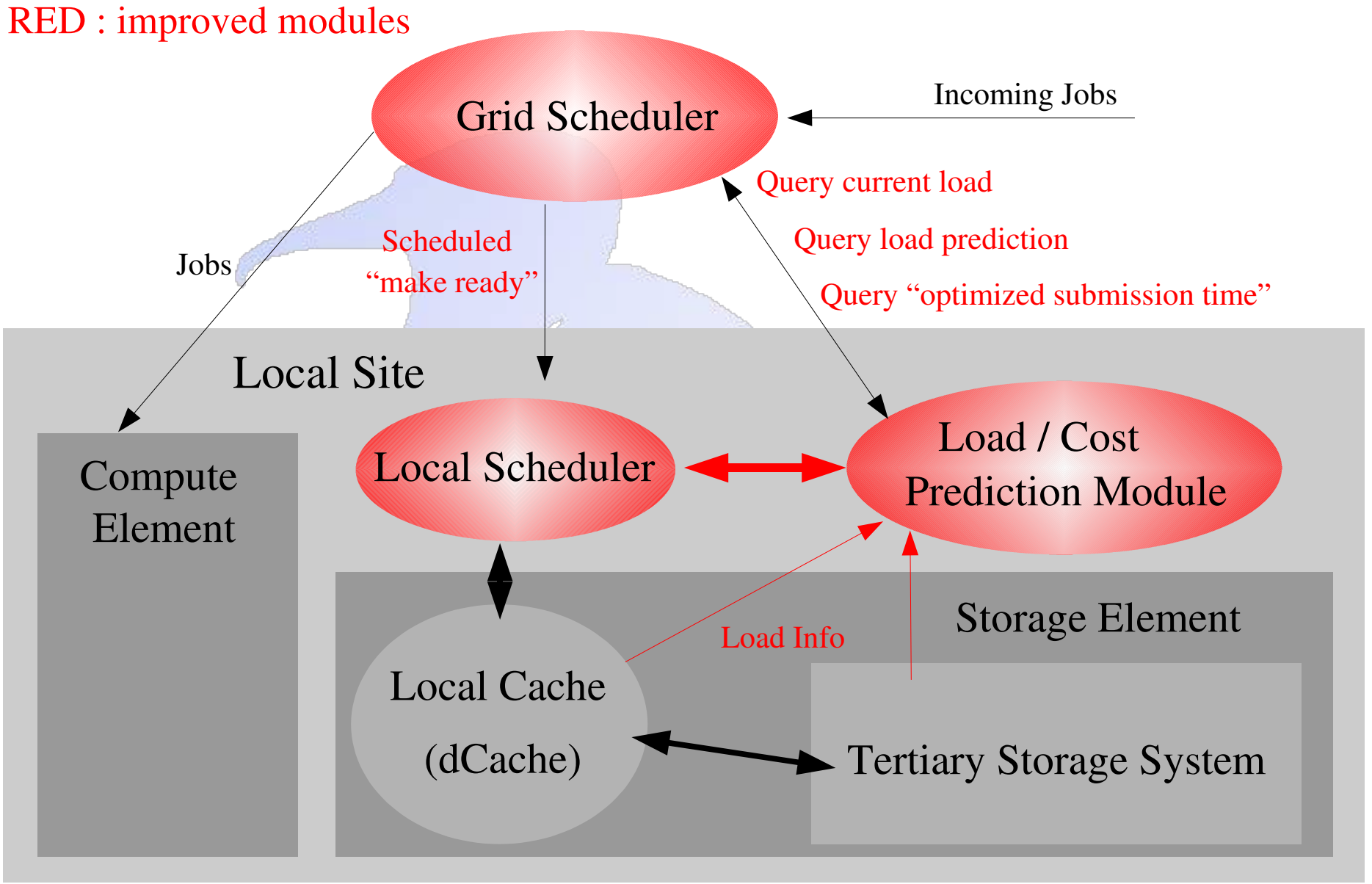
31 Tbytes disk space
244 Tbytes tape space

File Size vers. Data used

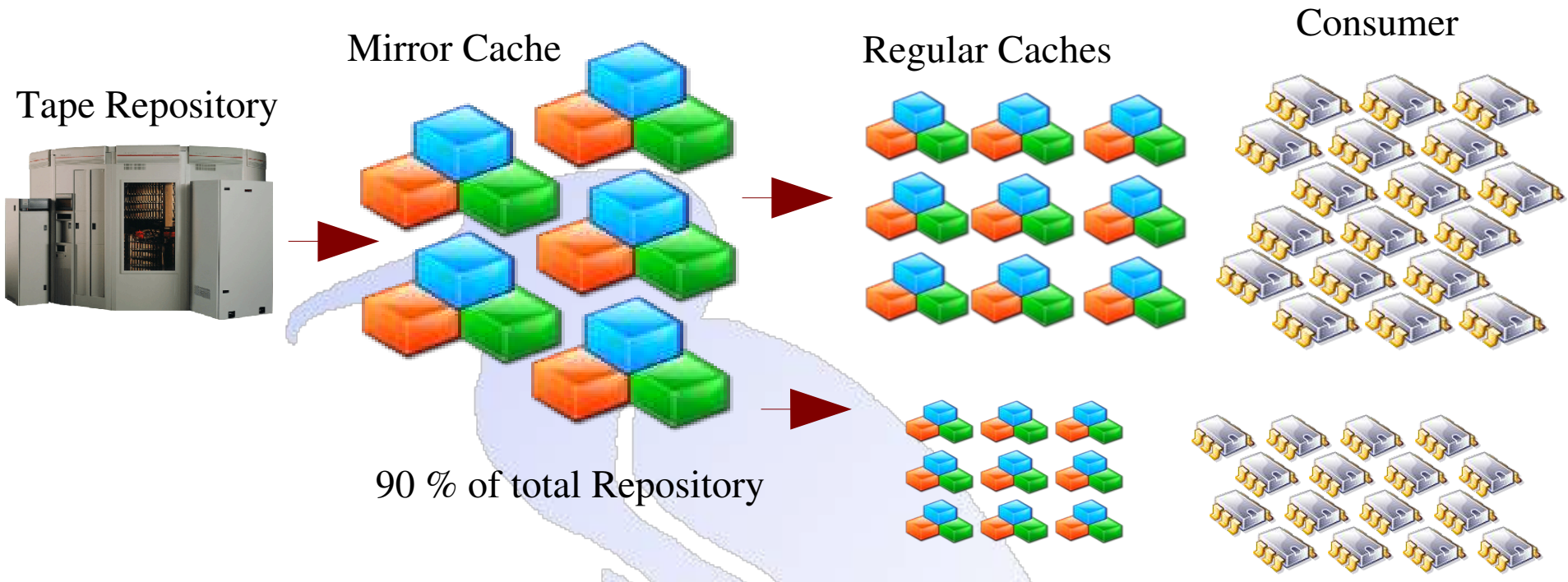


Restore vers. Transferred (H1 only)

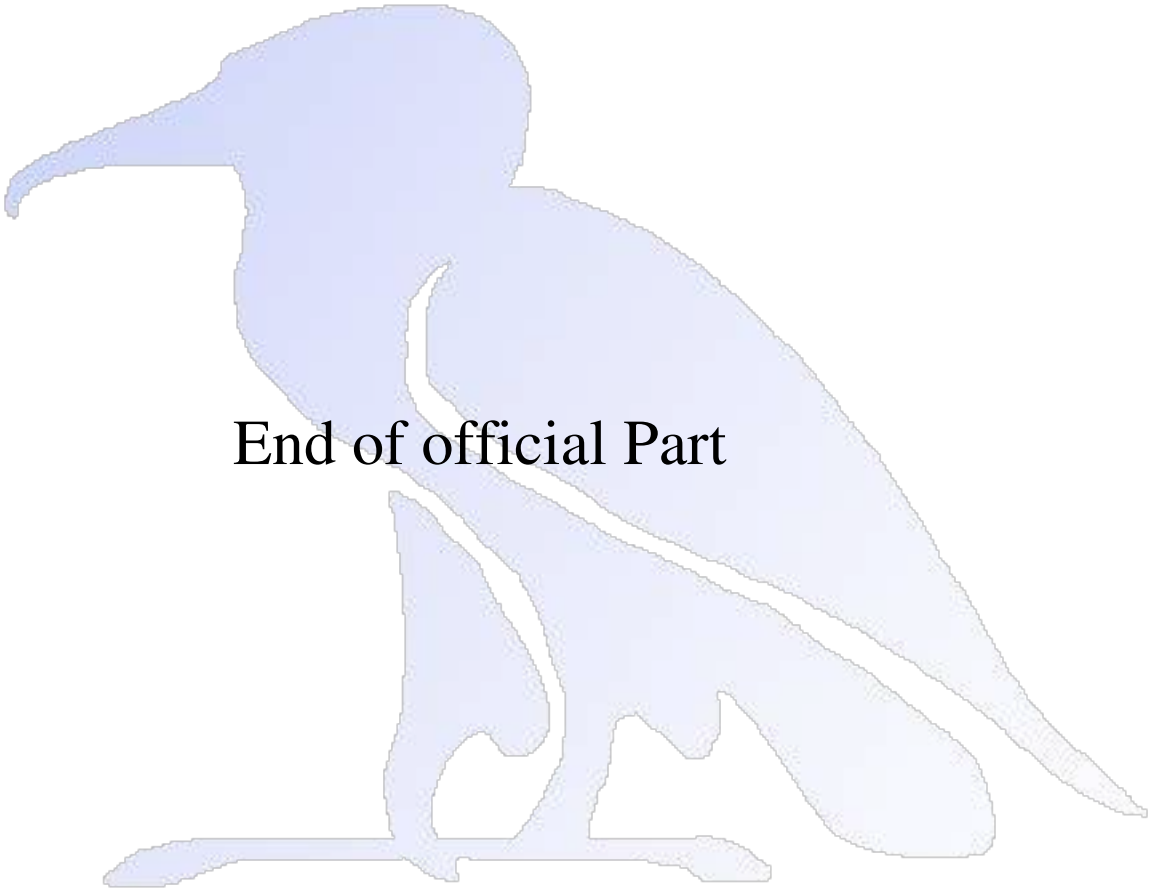



RED : improved modules
Grid Scheduler
Incoming Jobs
Query current load
Query load prediction
Query "optimized submission time"
Jobs
Scheduled
"make ready"
Local Site
Compute Element
Local Scheduler
Load / Cost Prediction Module
Storage Element
Load Info
Local Cache
(dCache)
Tertiary Storage System

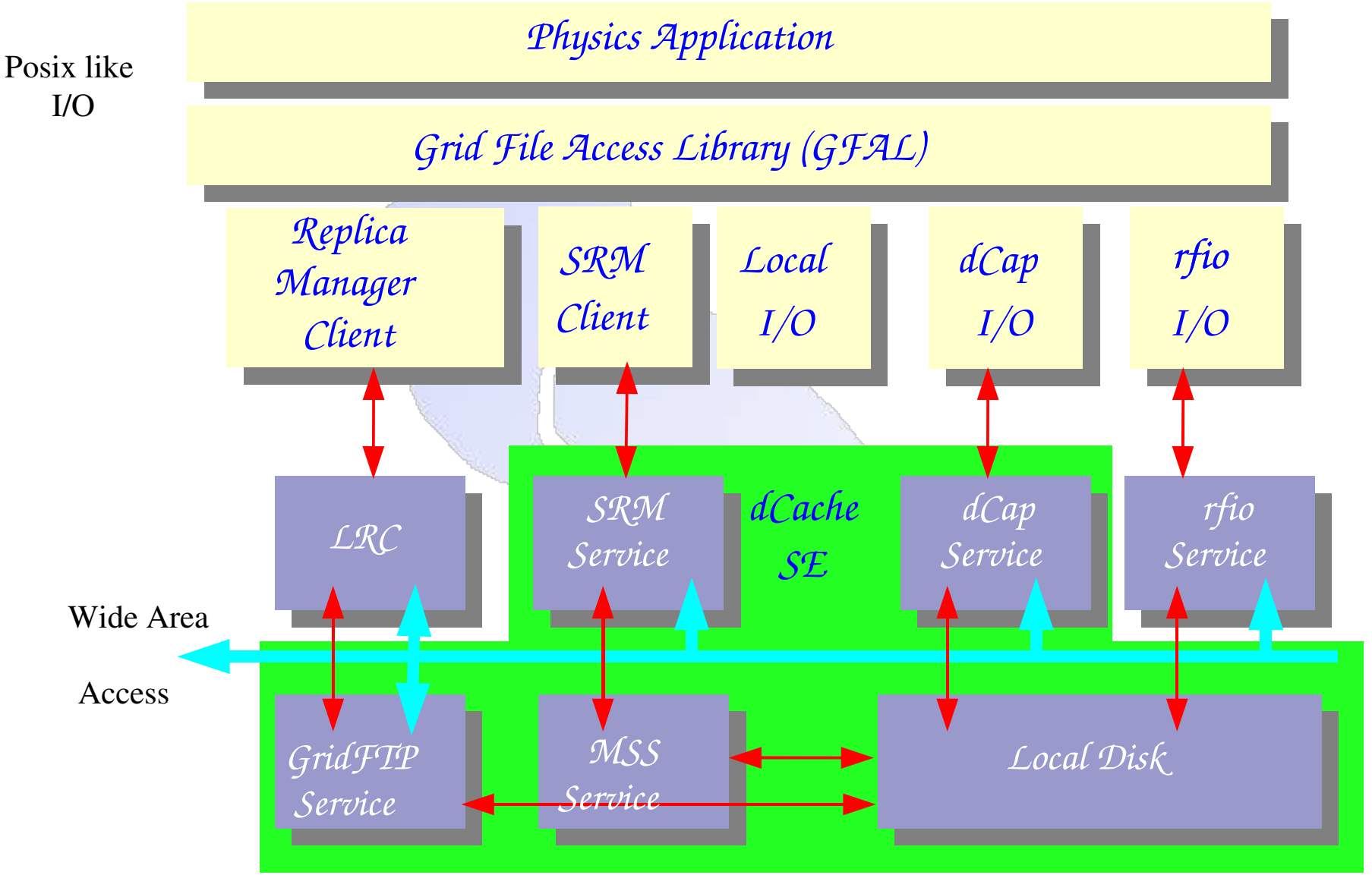
The Mirror Cache



- Nearly all Tape Data on *Mirror Cache*
- *Mirror Cache* has highest possible data density (lowest dollars/TBytes)
- Controlled number of high speed streams between *Mirror Cache* and *Regular Cache*
- *Mirror Cache* behaves like HSM (except for mount/dismount delays)
- *Mirror Cache* disks (or disk clusters) switched OFF if not accessed
- HSM to *Mirror Cache* transfers necessary only after disk replacement

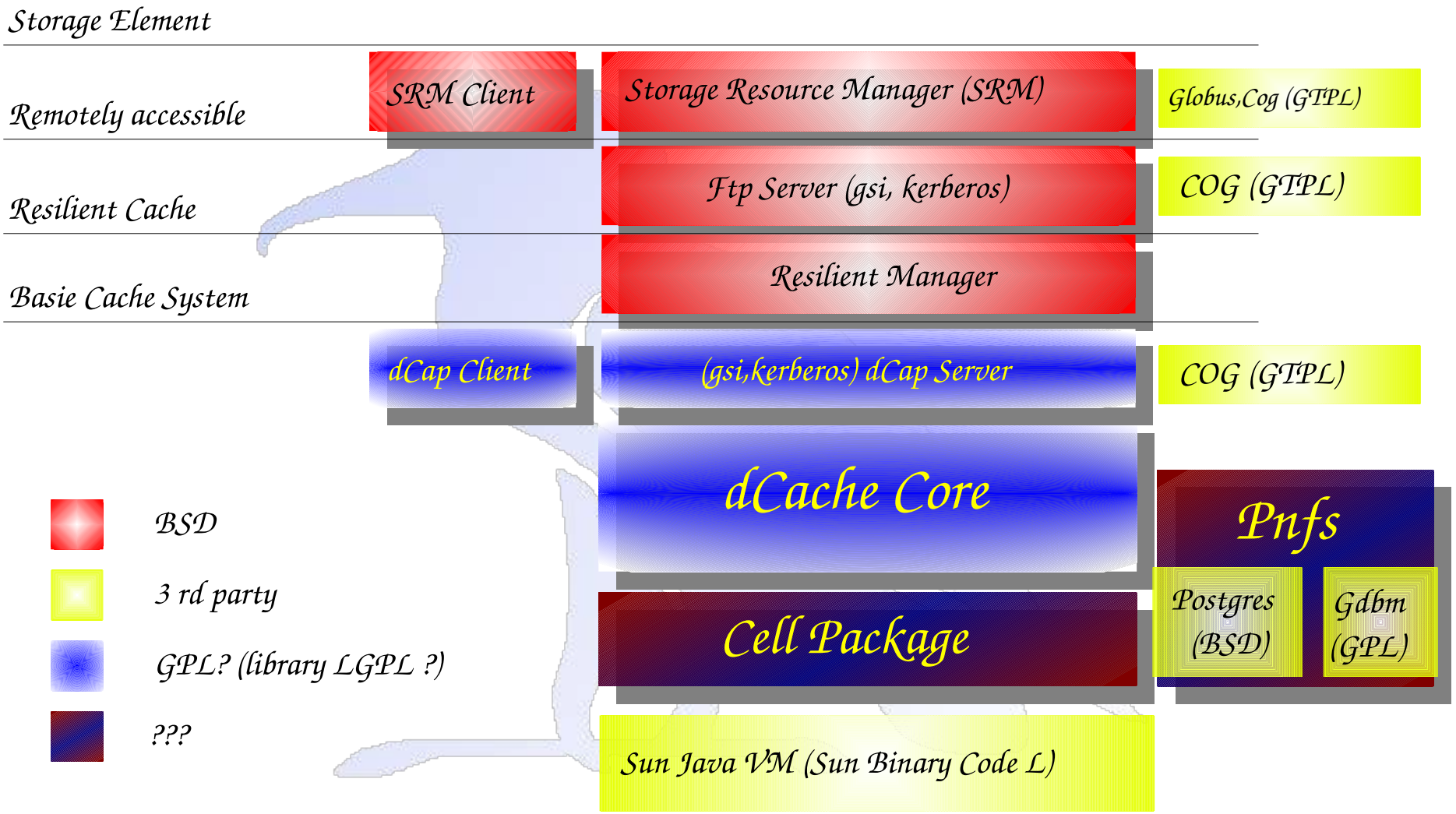


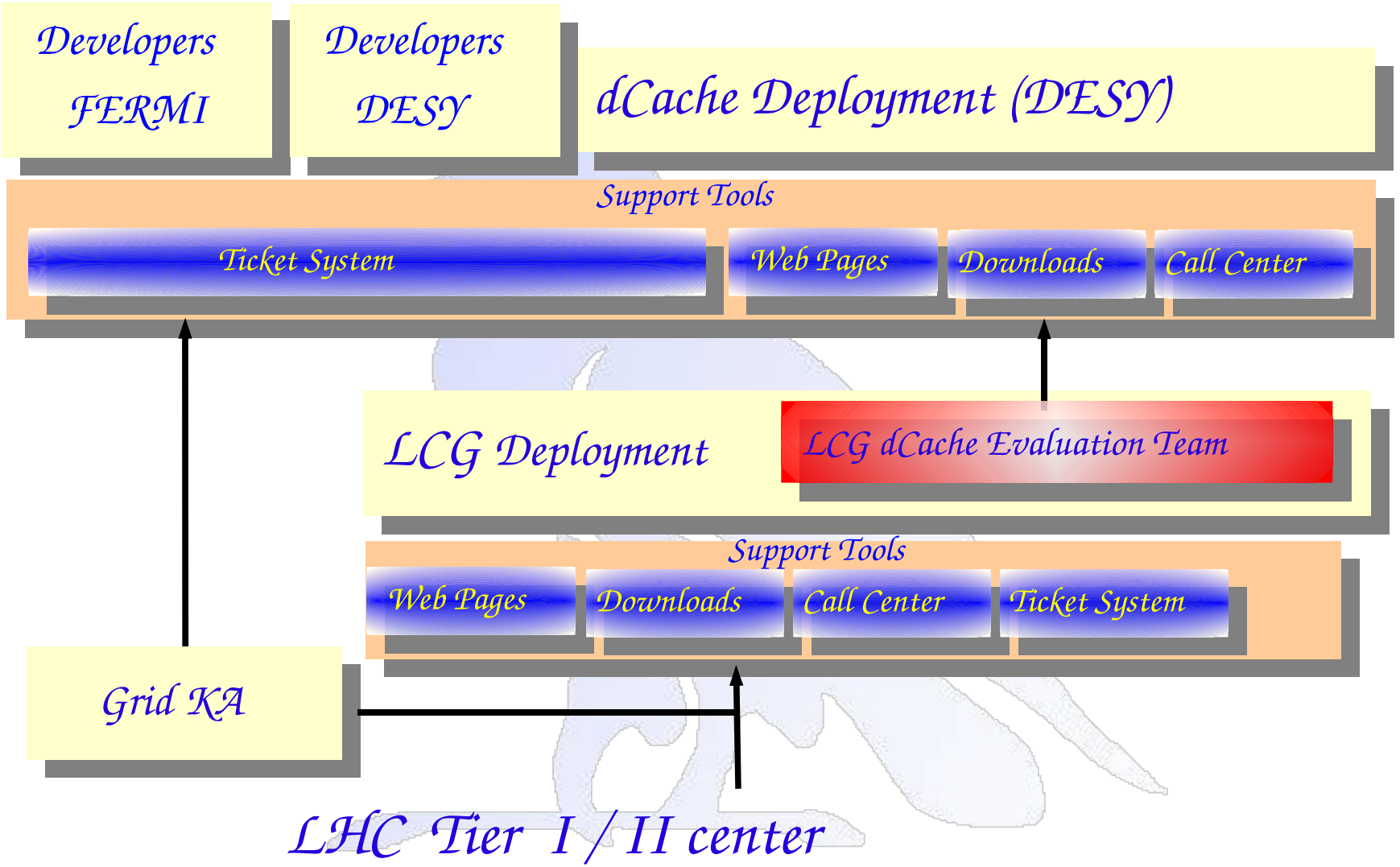
End of official Part



Source : Michael Ernst 18/5/2004

dCache Component License Model




Developers FERMI
Developers DESY
dCache Deployment (DESY)
Support Tools
Ticket System
Web Pages
Downloads
Call Center
LCG Deployment
LCG dCache Evaluation Team
Support Tools
Web Pages
Downloads
Call Center
Ticket System
Grid KA
LHC Tier I / II center