The dCache Storage Element

Patrick Fuhrmann

for the dCache people

WORKSHOP ON STATE-OF-THE-ART IN SCIENTIFIC AND PARALLEL COMPUTING

The Team

Responsibility, dCache

Patrick Fuhrmann Rob Kennedy

Responsibility, SRM

Timur Perelmutov

Core Team (Desy and Fermi)

Jon Bakken

Ted Hesselroth

Alex Kulyavtsev

Birgit Lewendel

Dmitri Litvintsev

Tigran Mrktchyan

Martin Radicke

Owen Synge

Vladimir Podstavkov

External Development

Nicolo Fioretti, BARI

Abhishek Singh Rana, SDSC

Support and Help

Maarten Lithmaath, CERN

Owen Synge, RAL, gridPP

Fermilab

dCache

DESY

SciDAC

the globus alliance

Grid Ftp

SRM - Collaboration

SRM v2.2

UCSD

gPlazma

Grid based authorization

D-GRID

HEPCG Project

Scalable Storage Element

Coscheduling

CEI Computer Engineering Institute

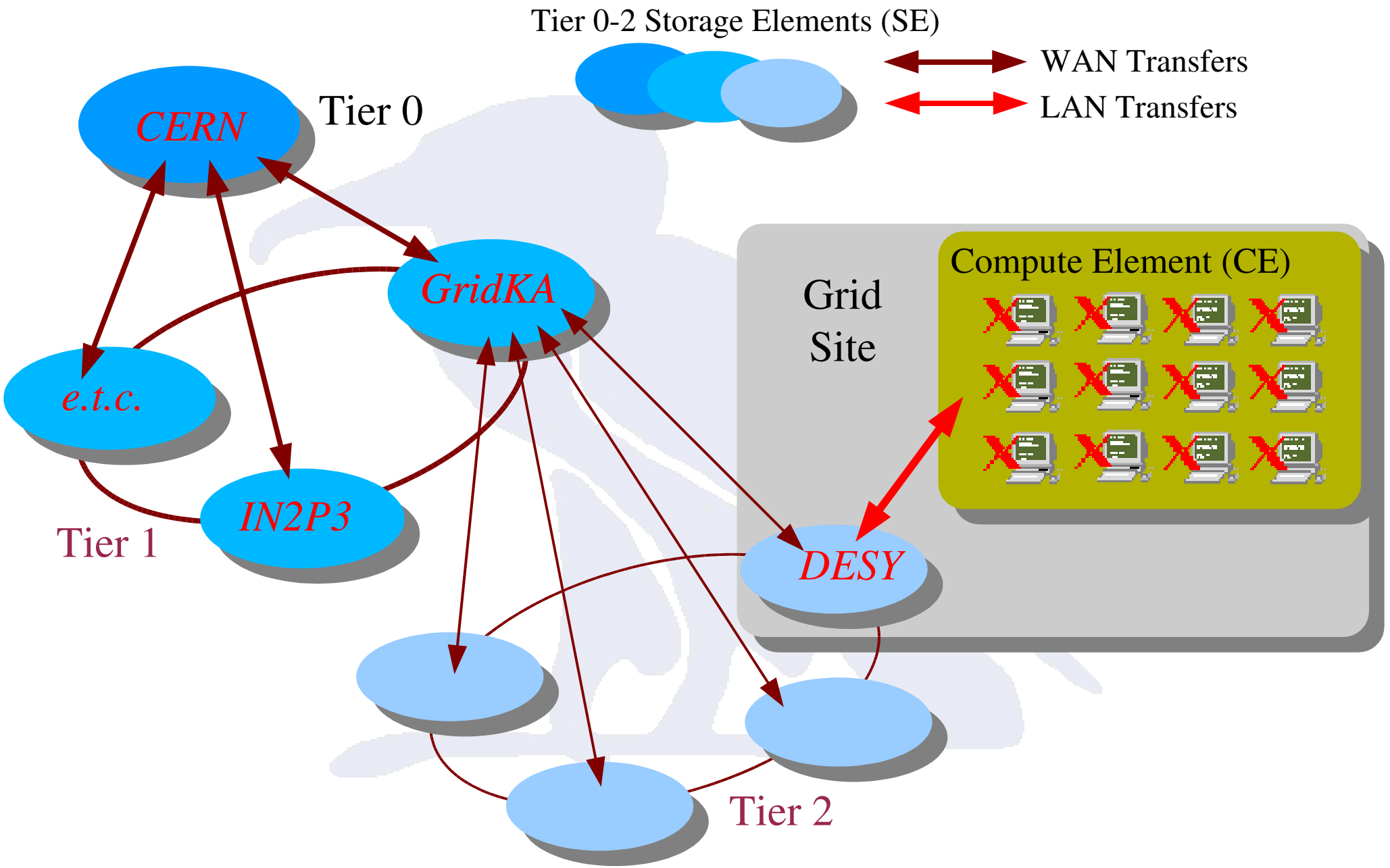
Integration Project (DGI)

Core Grid Middleware

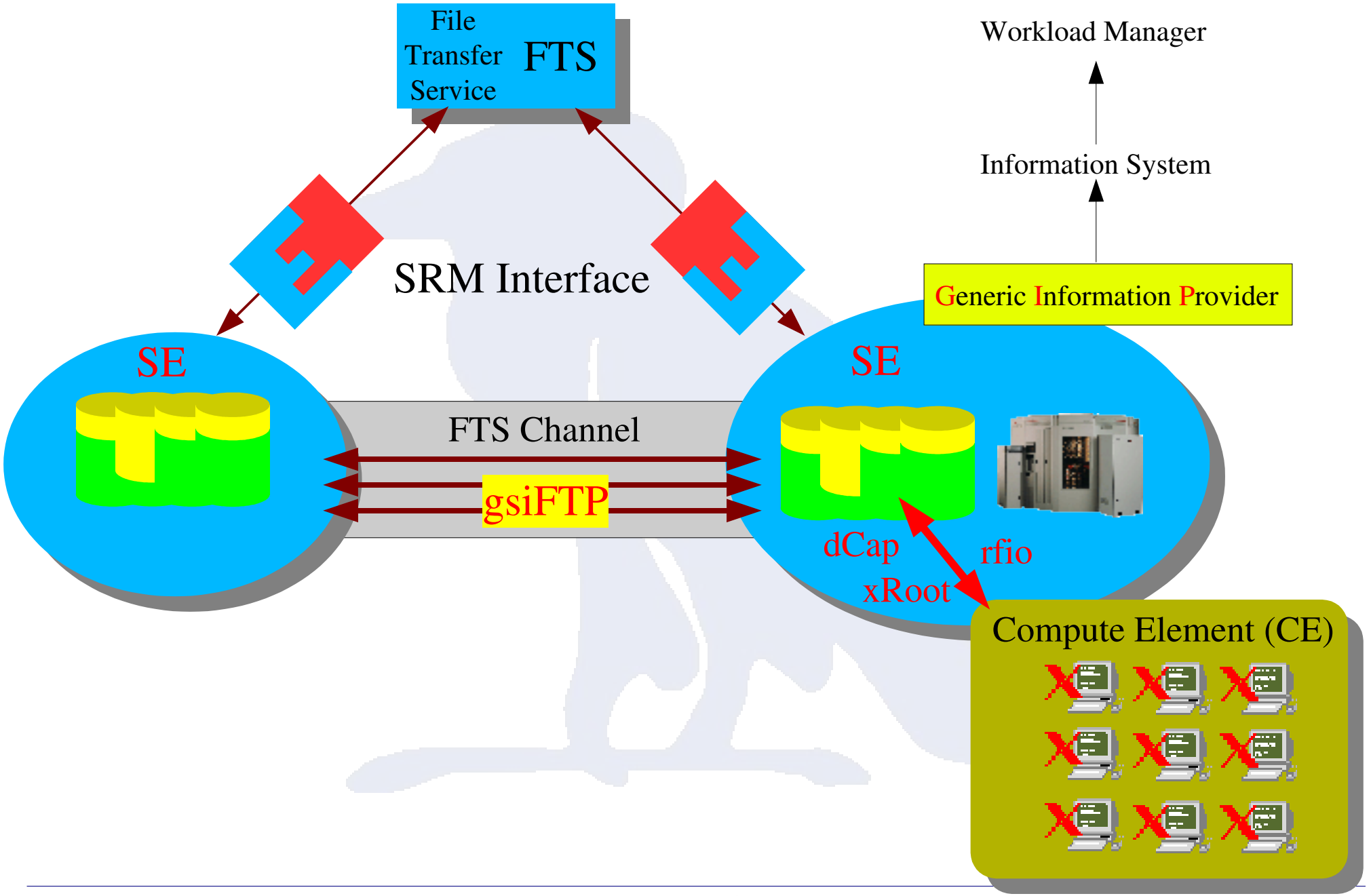
@ Jülich (FZJ/ZAM)

Fast Zoom into

Grid Enabled Managed Storage

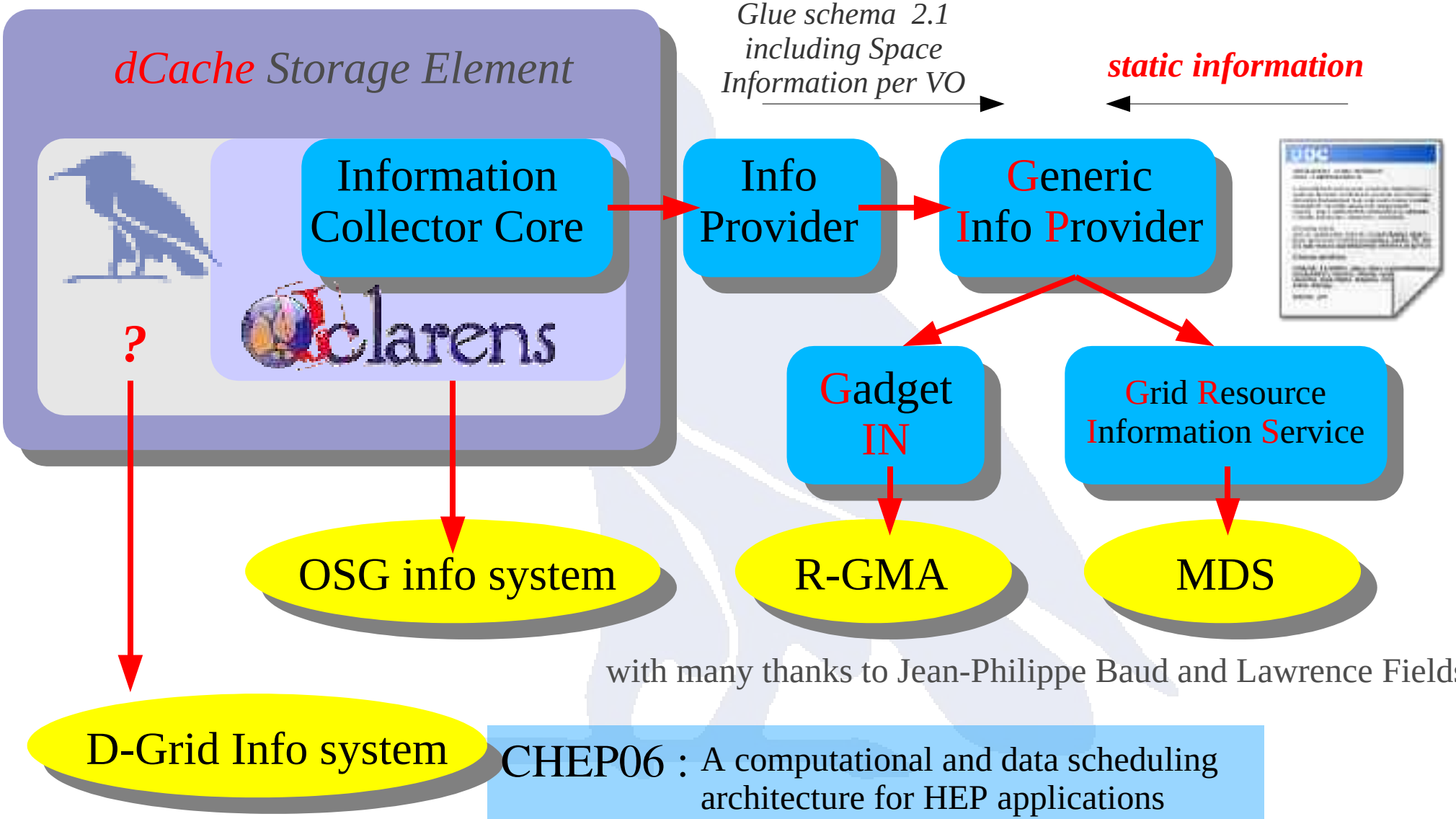

Tier 0-2 Storage Elements (SE)
WAN Transfers
LAN Transfers
CERN
Tier 0
GridKA
e.t.c.
IN2P3
Tier 1
Grid Site
Compute Element (CE)
DESY
Tier 2

SE and data transfers



Information Provider Details

by courtesy of Nicolò Fioretti



with many thanks to Jean-Philippe Baud and Lawrence Fields

CHEP06 : A computational and data scheduling architecture for HEP applications

by Martin Radicke and Lars Schley

SRM is :

From the SRM project pages (sdm.lbl.gov/srm-wg/index.html) :

This is an international collaboration among CERN,FNAL, JLAB, LBNL and RAL.

From the SRM fermi pages (srm.fnal.gov) :

SRM is the Storage Resource Manager layer providing storage and location independent access to data.

Technically :

- *Prepares for data transfer (not transfer itself) by storage URL (SRUL)*
- *Negotiates data transfer protocol (theoretically).*
- *May initiate restore of data from back-end storage systems.*
- *Delivers 'transfer url' (TURL) for subsequent transfer (gsiFtp,httpg).*
- *Supports directory functions including file listings.*
- *Supports space reservation functionality (implicit and explicit via space tokens)*
- *Supports 'property spaces' :*

File Properties resp. Property classes

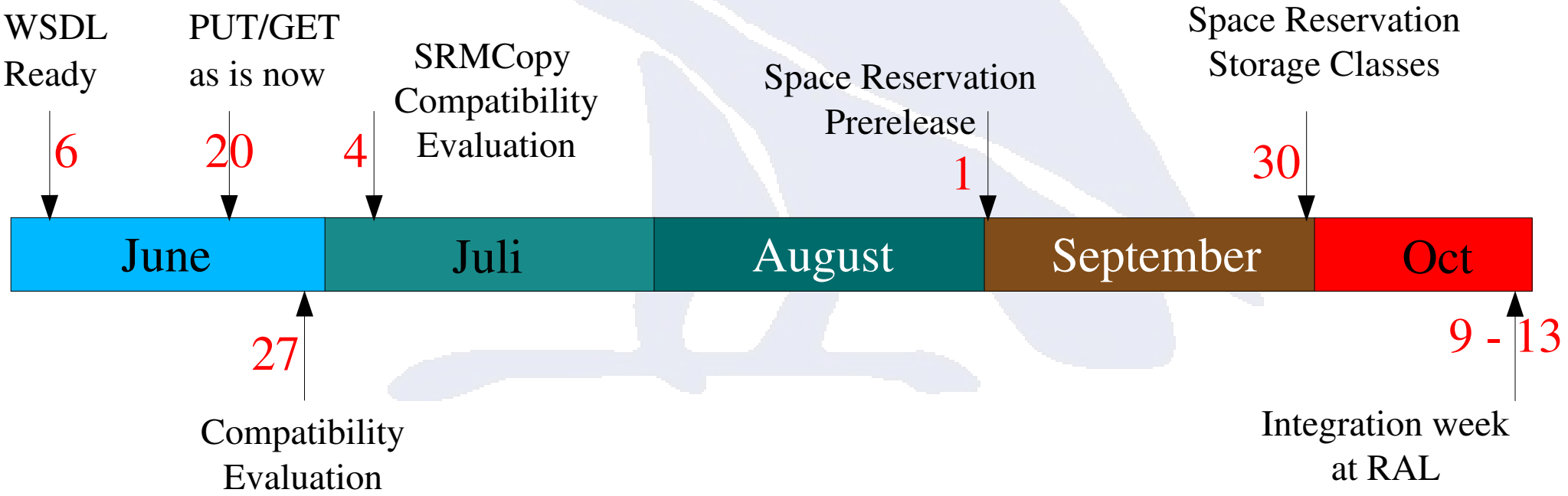
Media Quality	Persistence	Availability
probability of data loss	permanent *** volatile	how long does it take to get this file ready for I/O

Tape0 Disk1
Tape1 Disk0
Tape1 Disk1

SRM 1.X has been a great success

SRM 2.X is an unjustifiable huge amount of work for SE implementors.

SRM Interface implementation working group results (WLCG SRM definition V2.2)



Storage Elements currently available

CASTOR

Developed at CERN; CERN's main repository
HSM included
For huge installations only
Support available but requires man power compensation to CERN

dCache

Developed at DESY/FERMI
can talk to many HSM's
From small to huge installations (> 200 Tbytes on disk)
Support for free (support@dCache.ORG)

DPM Disk Pool Manager

Developed at CERN
for HSM less installations only
From small to medium installations
Support not clear

Berkeley DRM

Developed and supported by LBL (Lawrence Berkeley National Laboratory)

Used by Open science grid (OSG)

Part of VDT (Virtual Data Toolkit, Globus)

NORDUGRID

ARC (Advanced Resource Connector & Smart Storage Element)

Non Storage Elements

StoRM

Storage Resource Manager

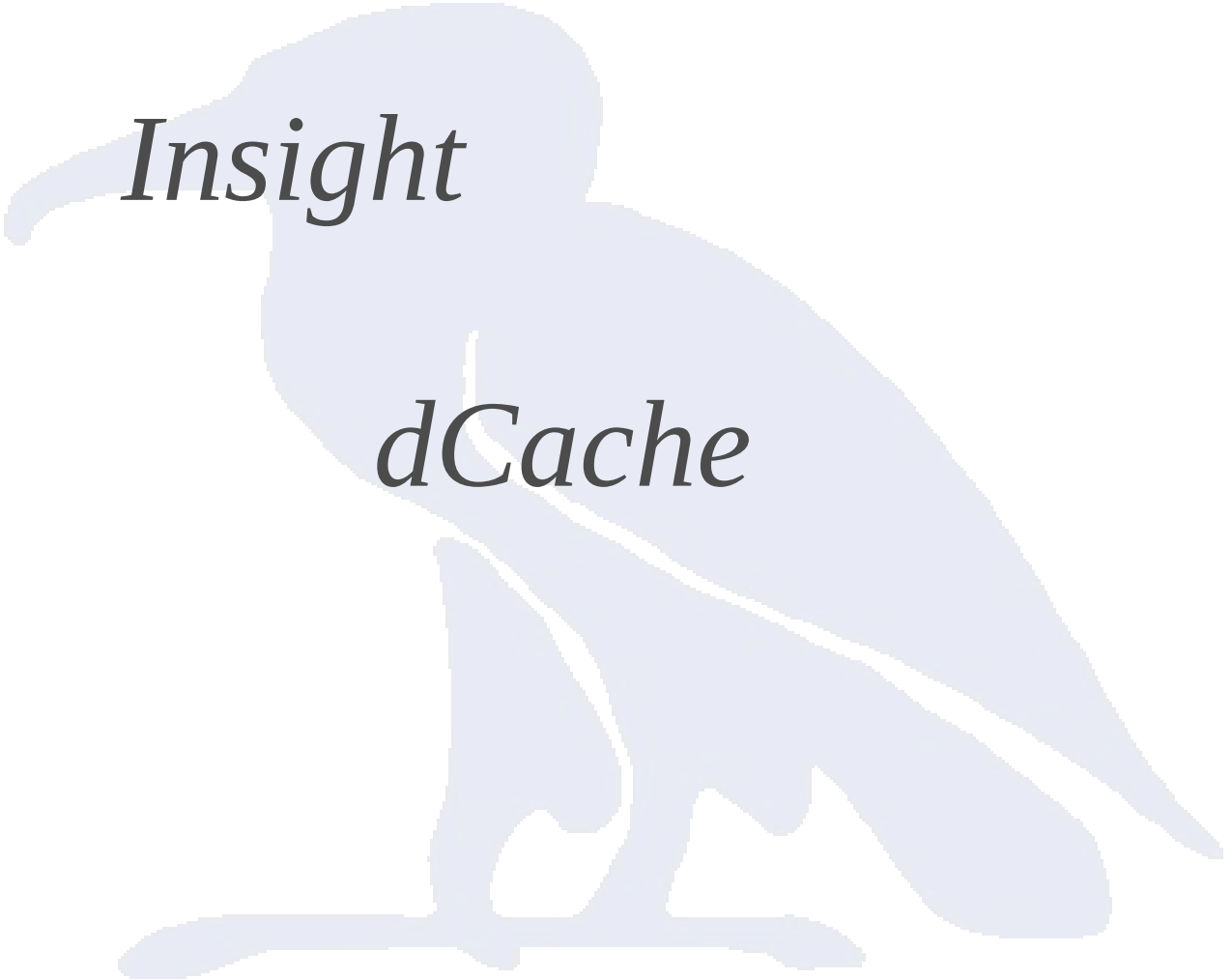
StoRM

Developed at INFN, CNAF, Italy
Independent SRM implementation
Interacts with regular filesystem
Enhanced support for GPFS (space reservation)

---

xRootD

Developed at SLAC
Mainly for analysis (fast opening of huge amount of files)



Insight

dCache

dCache is Managed Storage

Distributed Peta Byte Disk Store with single rooted file-system providing posix like and wide area access protocols.

Distributed cache system to optimize access to Tertiary Storage Systems

Grid Storage Element coming with standard data access protocols, Information Provider Protocols and Storage Resource Manager.

Basic Specification

Single 'rooted' file system name space tree

File system names space view available through an nfs2/3 interface

Data is distributed among a huge amount of disk servers.

Supports multiple internal and external copies of a single file

Supports 'posix like' access (dCap, xRoot) as well as various FTP dialects, (http) and the Storage Resource Manager Protocol.

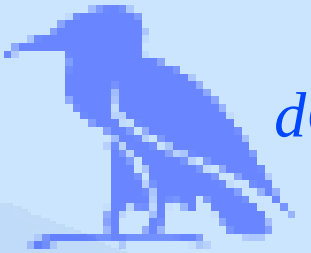
SRM for storage management

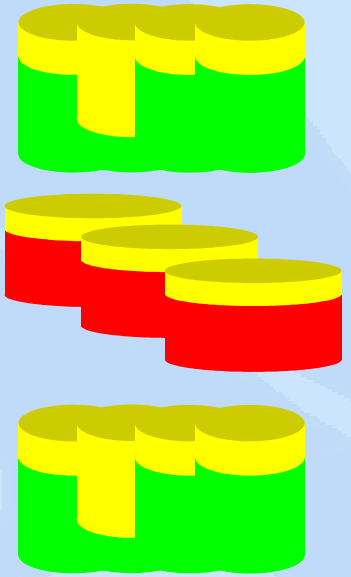
nfs2/3 for name space

/pnfs/<site>/<VO>/...

dCap, xRoot for random LAN

gsiFtp, http(g) for random WAN





Osm, Enstore,
Tsm, Hpss, ...

Selected Topics

File hopping I

- Automatic data set replication on hot spot detection.
- File replication on client read request (pools disallowed for reading)
- Dataset replication on arriving of datasets.

Collect and flush

Write Only Cache

Forward on
read
or on
arrival

Read Only Cache

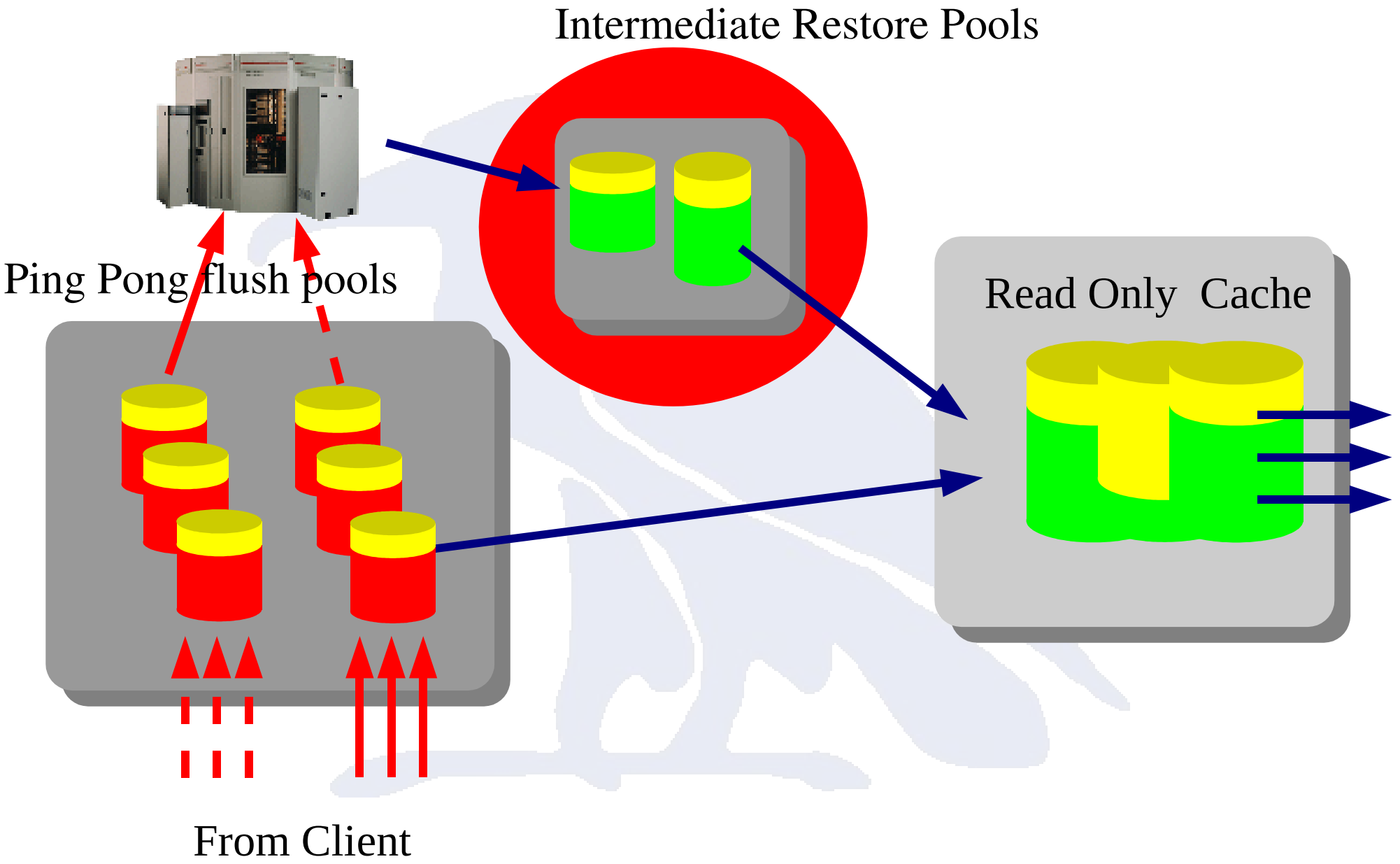
Replicate
on high load

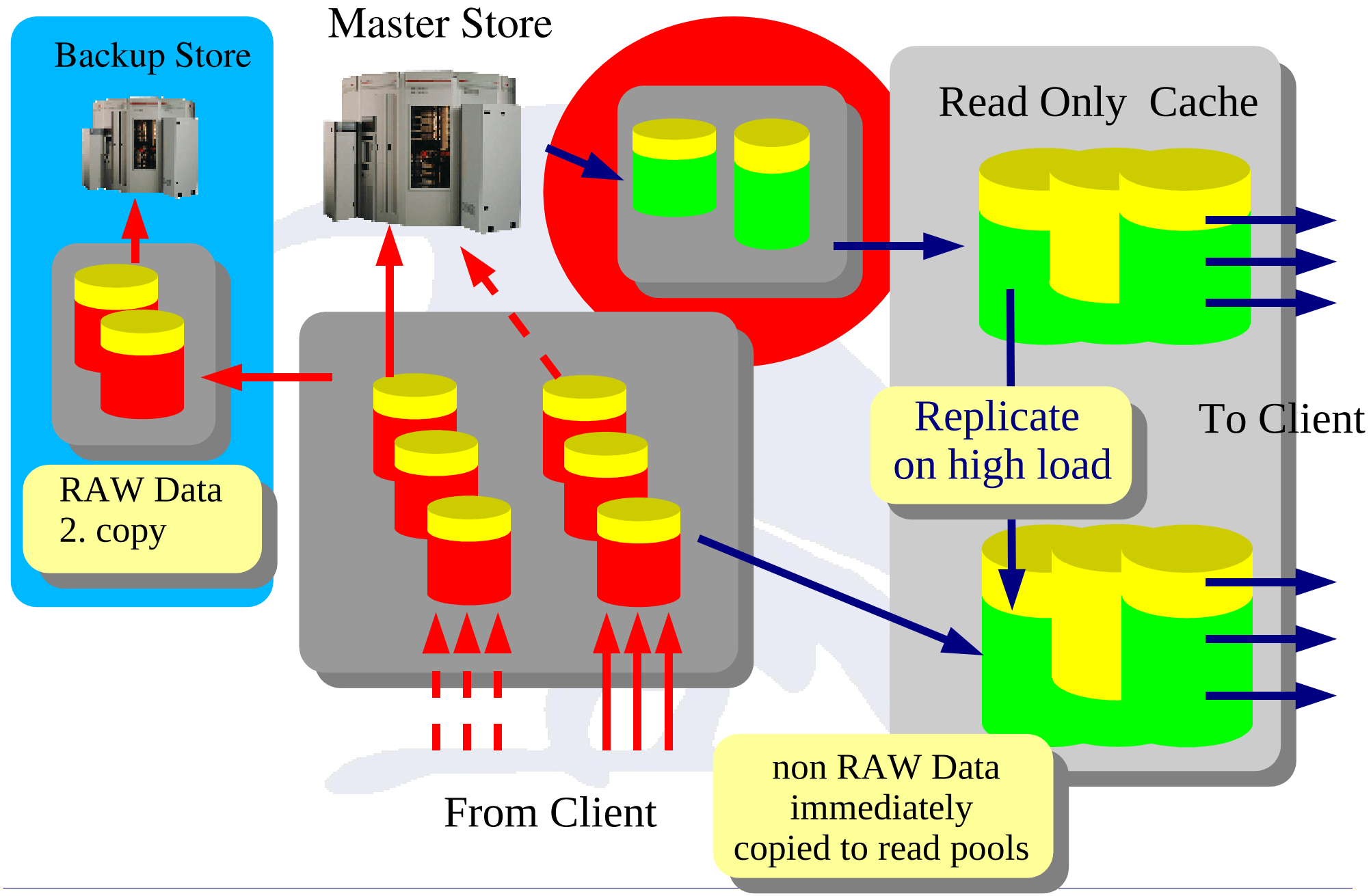
To Client

From Client

HSM interactions

- Datasets collected in write pools and flushed according to rules.
- Centrally controlled (Smart) flushing -> ping pong
- Datasets restored if requested but no longer in cache.
- Intermediate restore pool for HSM optimization.


Intermediate Restore Pools
Ping Pong flush pools
Read Only Cache
From Client


Backup Store
Master Store
Read Only Cache
RAW Data
2. copy
Replicate
on high load
To Client
From Client
non RAW Data
immediately
copied to read pools

Please refer to poster by Alexander Kulyavtsev

Resilient dCache (pools on worker nodes)

- Controls number of copies for each dataset in dCache
- Makes sure n < copies < m
- Adjusts replica count on pool failures
- Adjusts replica count on scheduled pool maintenance
- Makes use of local disk space when running on farm nodes
- doesn't work with HSM back-end yet

Improvements

- File copy operations (pool to pool) will be controlled by the PoolManager
- Pool Manager rules are honored (including : don't copy to same host/store)
- Pool Manager cost metrics is honored

And not to forget ...

- Destination pool selection by IP, directory, protocol, I/O direction.
- Final pool selection by space cost and pool node load.
- dCache instance partitioning.
- Extended proxy (certificate) support (OSG and LCG)
- Draining of pools for maintenance.
- Rich command line interface (via ssh).
- First version of GUI for admin and cpu/space cost analysis.
- Highly improved file system emulation (chimera) in evaluation phase.
- See 'dCache, the Book' for details.

Tier I centers :

- FNAl
- BNL
- IN2P3
- SARA
- Triump
- Nordu grid
- gridKa
- (still RAL)

Tier II centers :

Germany

LCG : Aachen, DESY, Freiburg, Dortmund,Darmstadt(GSI)

d-Grid : Juelich(ZAM), Berlin(ZIB)

UK

30 % of gridPP, UK

US

CMS : 7 sites, ATLAS 7 sites in preparation

Italy

INFN : Bari, Torino

Poland, Bulgaria, Spain

Canada, Taiwan

dCache, the Book

www.dCache.ORG

need specific help for you installation or help
in designing your dCache instance.

support@dCache.ORG

dCache user forum

user-forum@dCache.ORG