



dCache

a storage system for the future

Patrick Fuhrmann

Volker Gölzow

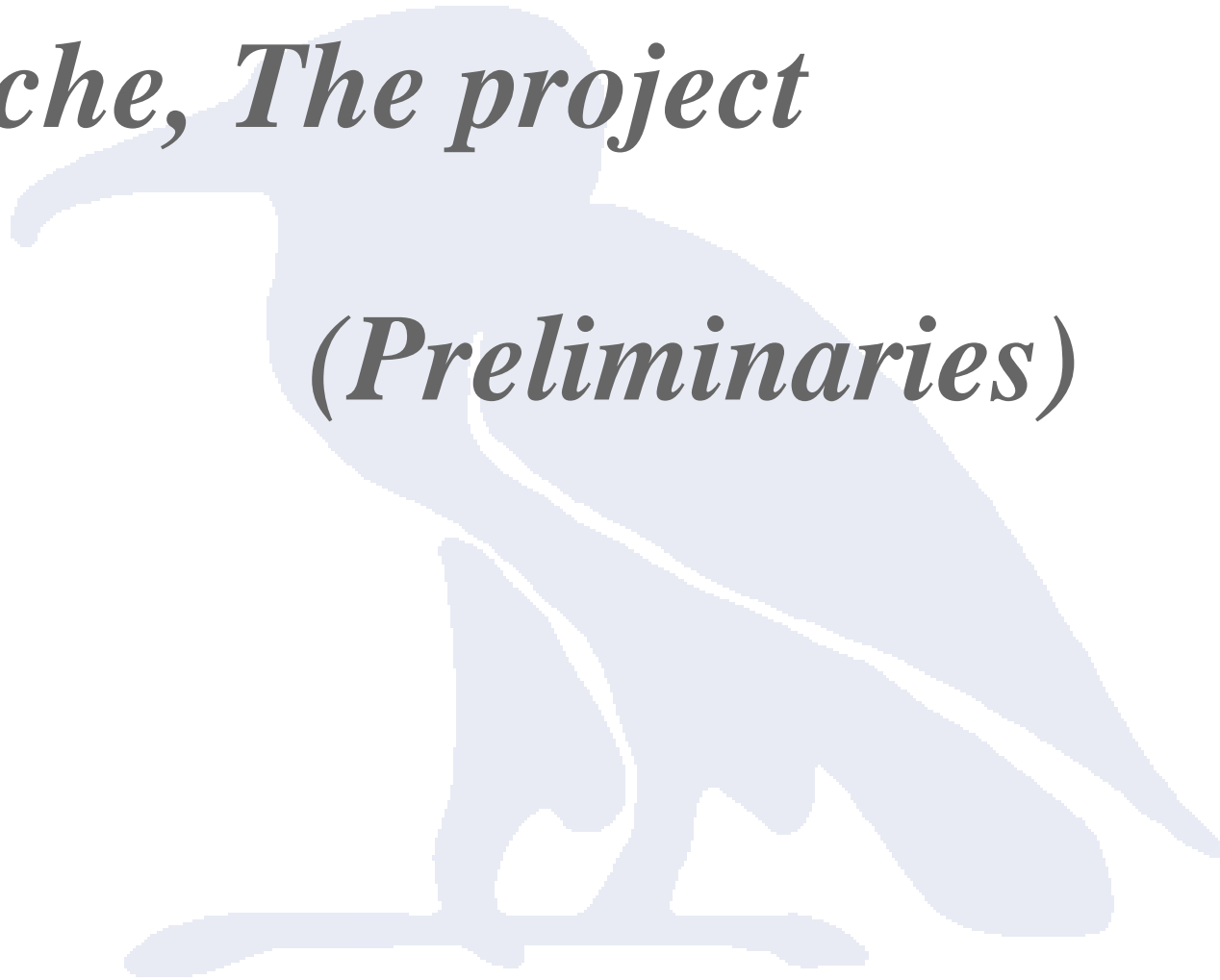
for the dCache people

Euro-Par 2006 : European Conference On Parallel Computing



dCache, The project

(Preliminaries)





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 Syngé

Vladimir Podstavkov

External

Development

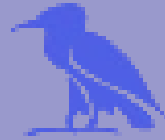
Nicolo Fioretti, BARI, Italy

Abhishek Singh Rana, SDSC, US

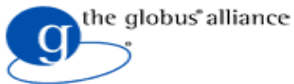
Support and Help

Maarten Lithmaath, CERN

N.N, CERN



dCache

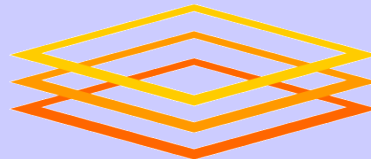


Grid Ftp

SRM -
Collaboration
SRM v2.2



30 % of all
LCG SE's



Open Science Grid
Mainly the large SE's
in OSG



gPlazma
Grid based authorization



HEPCG Project

Scalable Storage Element

Coscheduling



Integration Project (DGI)

Core Grid Middleware
@ Jülich (FZJ/ZAM)



*Fast Zoom into **LCG***

Grid Enabled Managed Storage



27 Km super conducting double ring

Proton collisions

20 events / collision

14 TeV center of mass

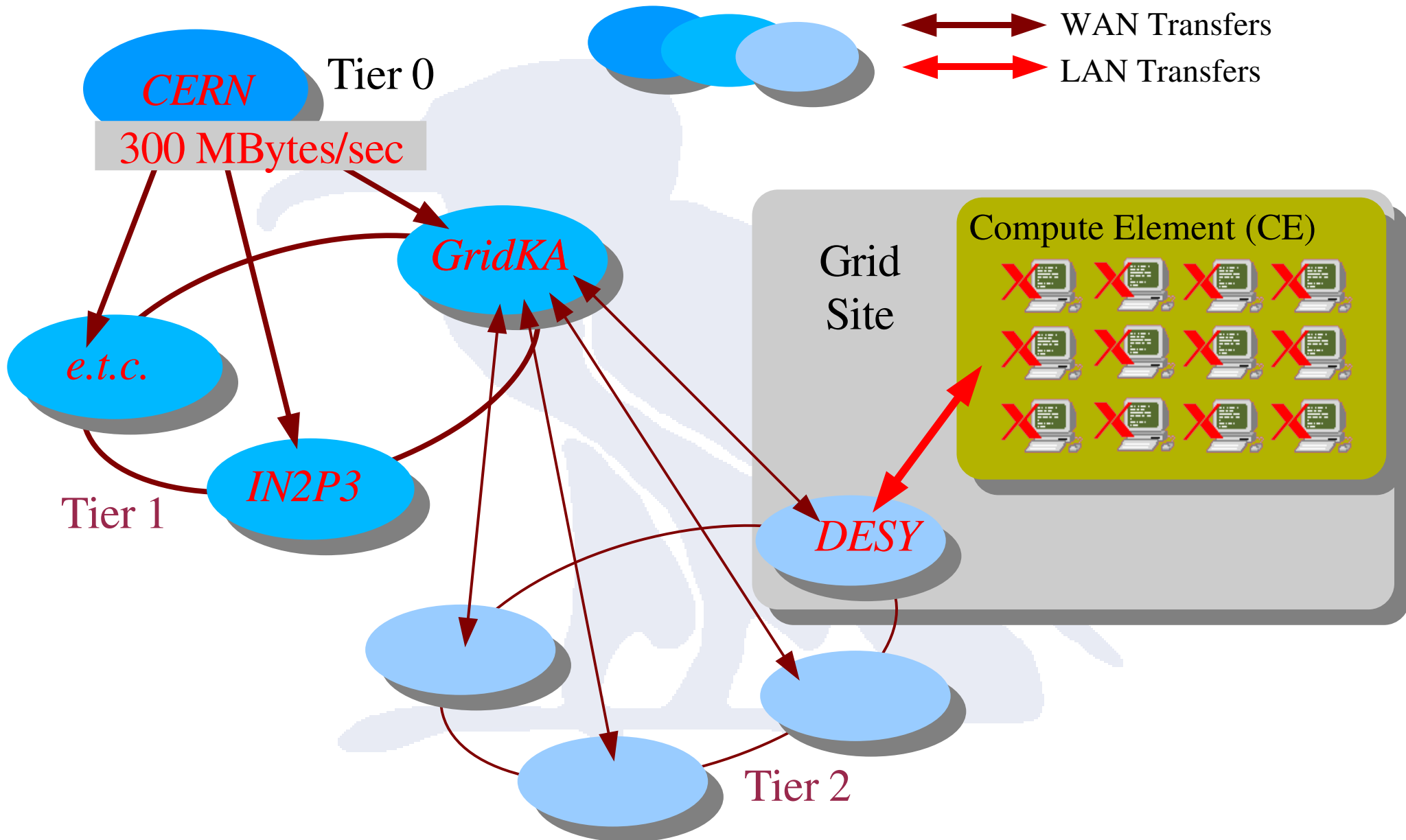
300 Mbytes/second raw data output

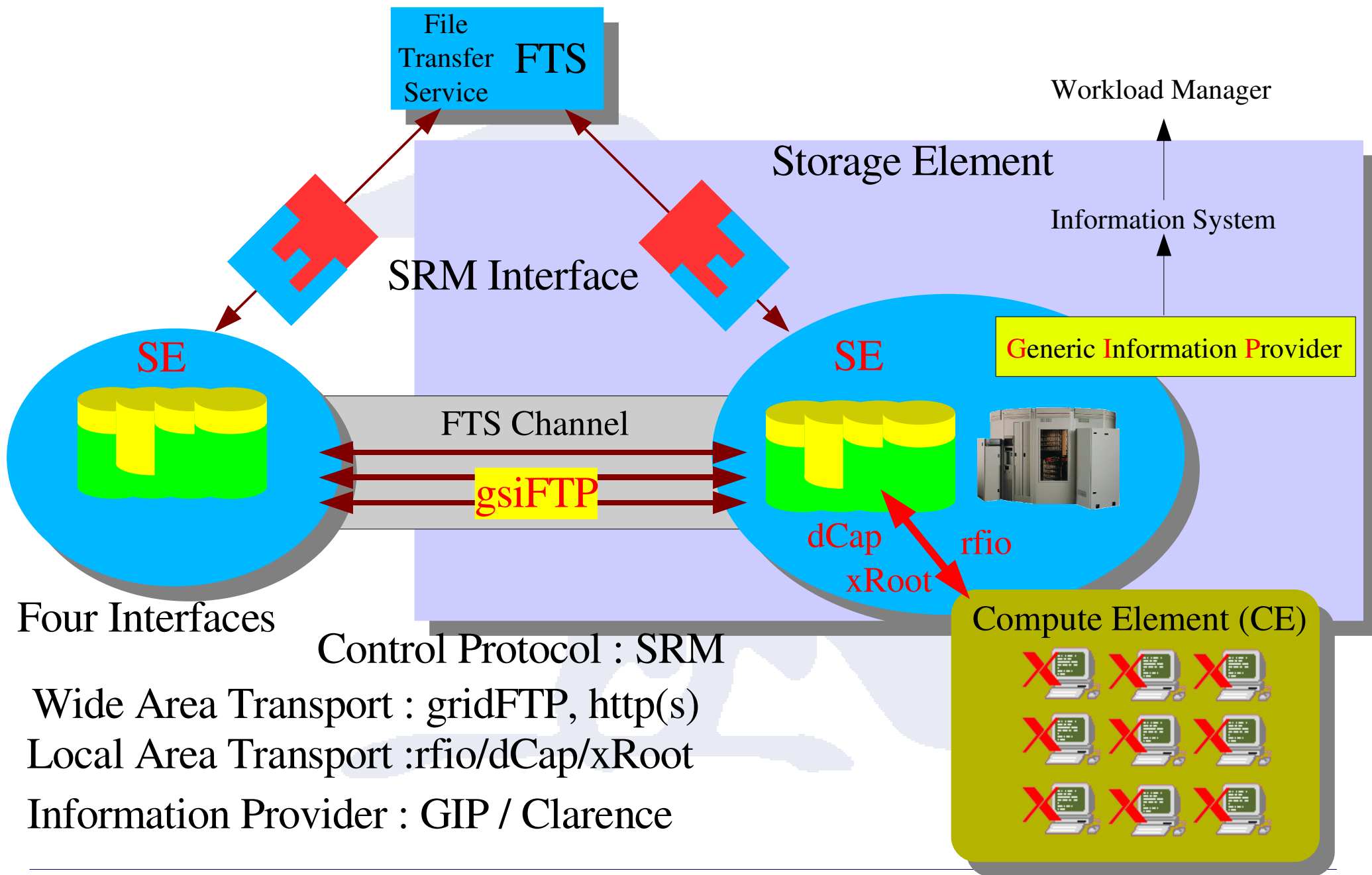
1 TeV is about the energy of motion of a flying mosquito. What makes the LHC so extraordinary is that it squeezes energy into a space about a million million times smaller than a mosquito. (Stolen from CERN Web Pages)





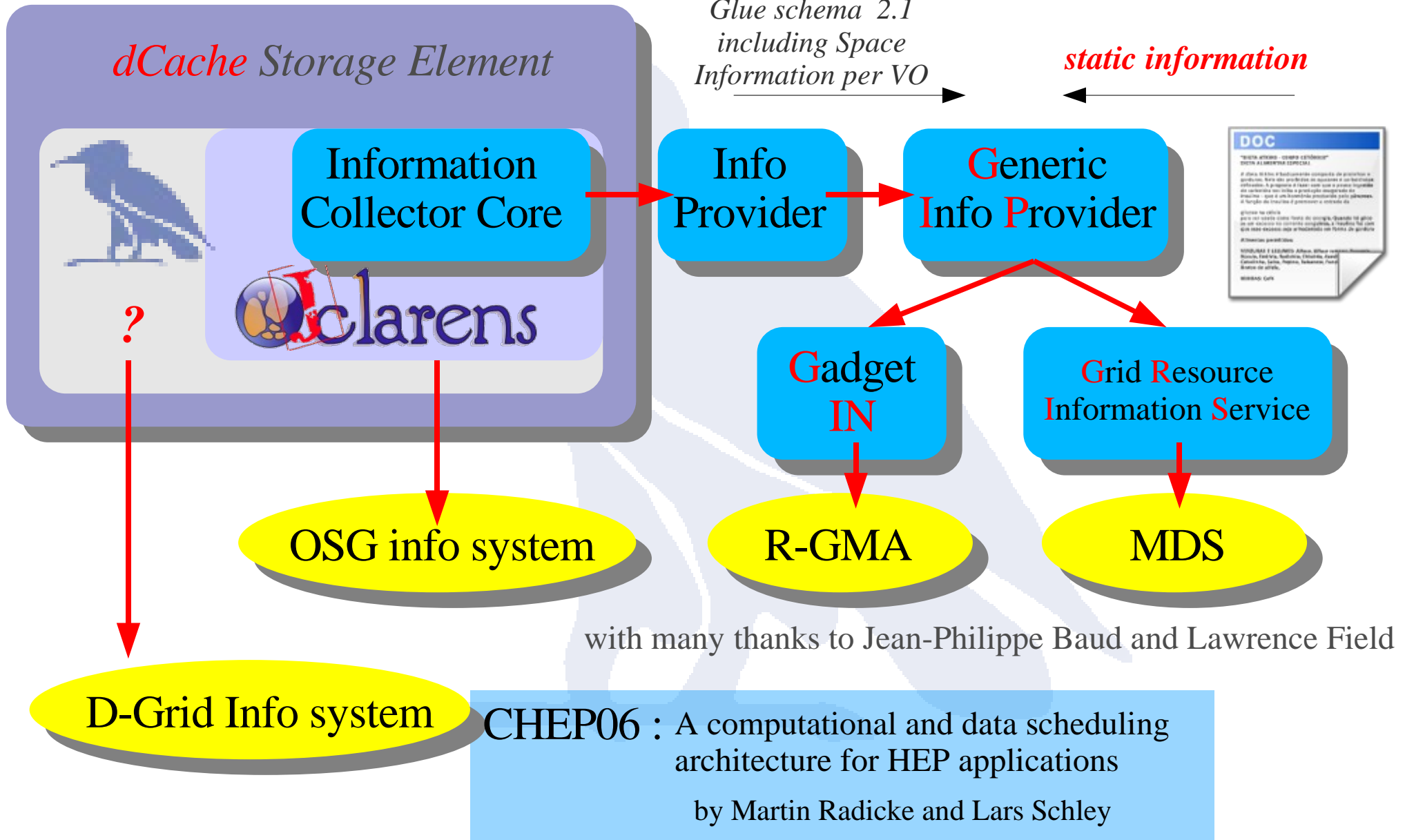
Tier 0-2 Storage Elements (SE)







by courtesy of Nicolò Fioretti





SRM is a Storage Management Abstraction

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

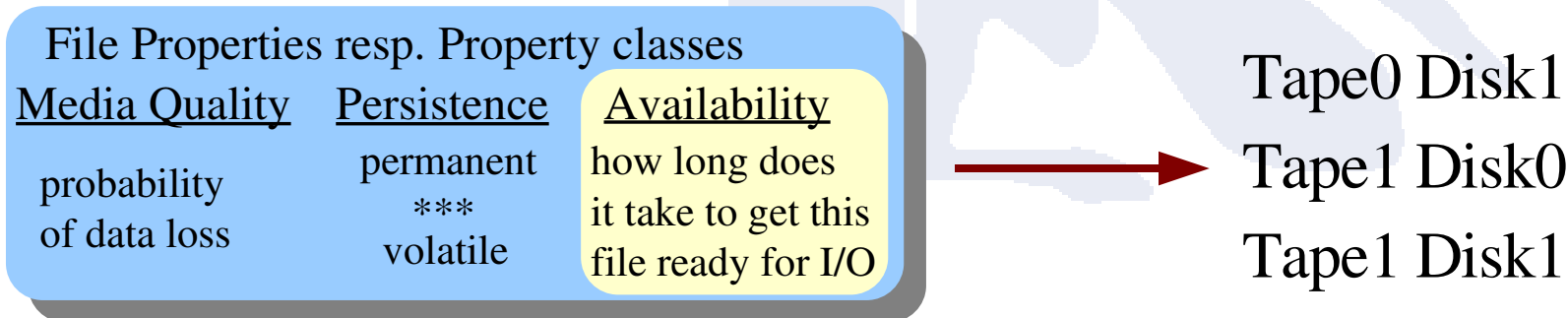
This is an international collaboration among CERN, FNAL, JLAB, LBNL and RAL.

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' :*





*Storage Elements
currently available*





CASTOR

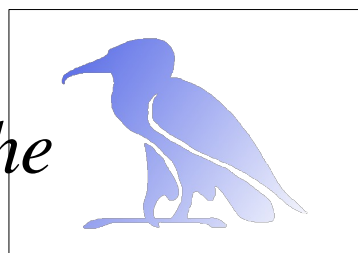


Developed at CERN; CERN's main repository

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)



ARC (**A**dvanced **R**esource **C**onector & **S**mart **S**torage **E**lement)

Non Storage Element (SRM Interface only)



StoRM

Developed at INFN, CNAF, Italy

Independent SRM implementation

Interacts with regular filesystem

Enhanced support for GPFS (space reservation)





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

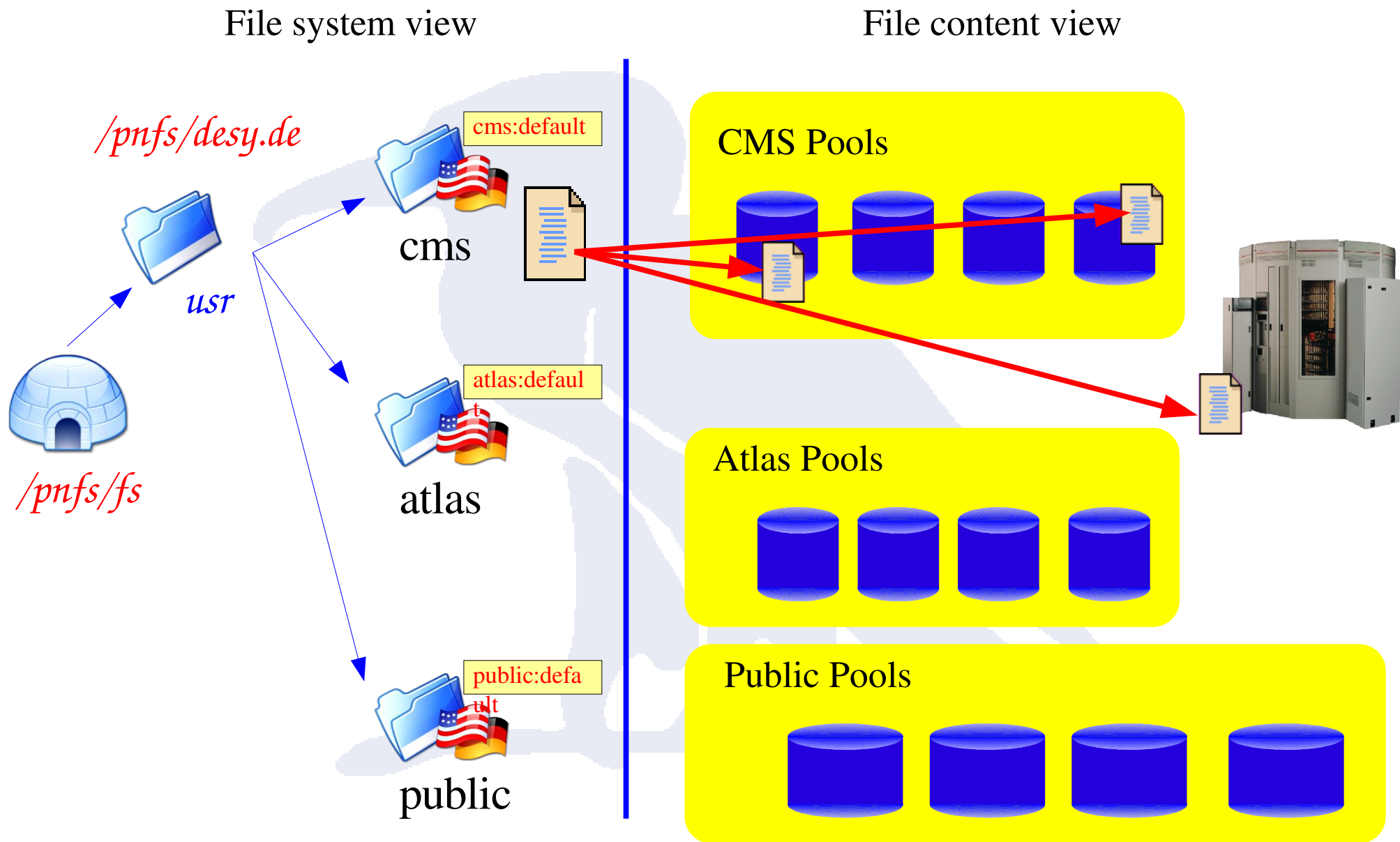
Strict separation between data and name-space

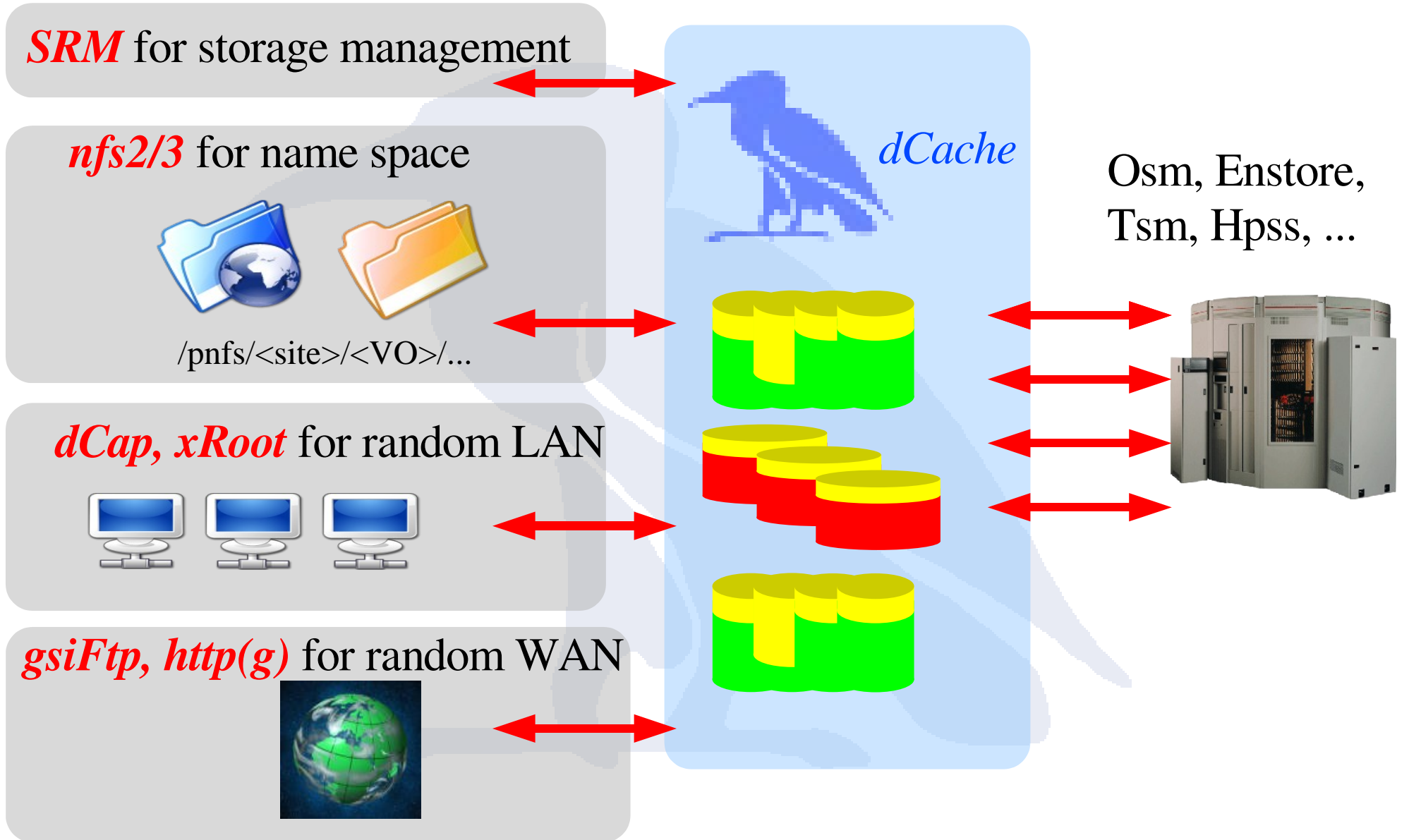
File system names space view available through an nfs2/3 interface

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.



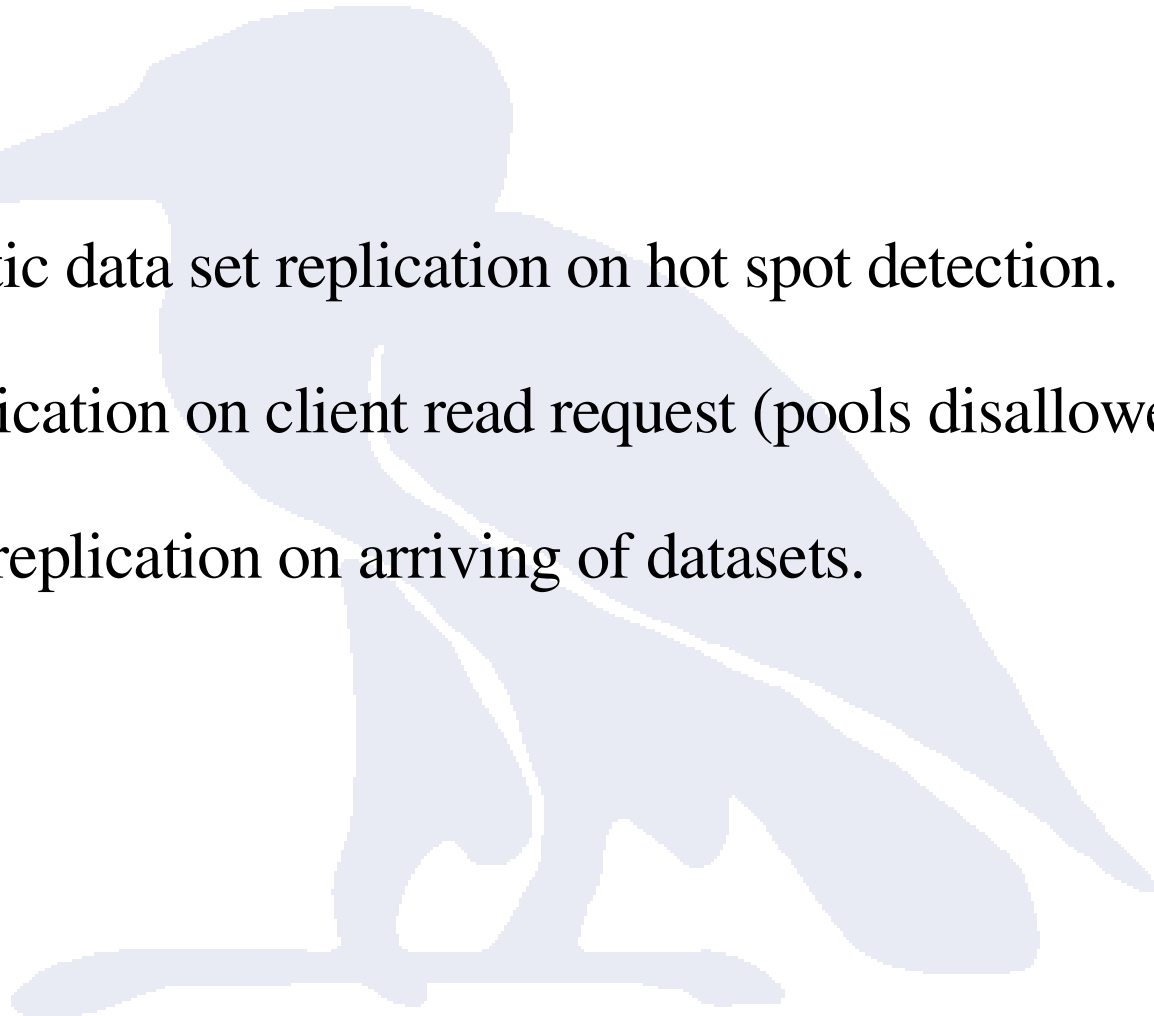


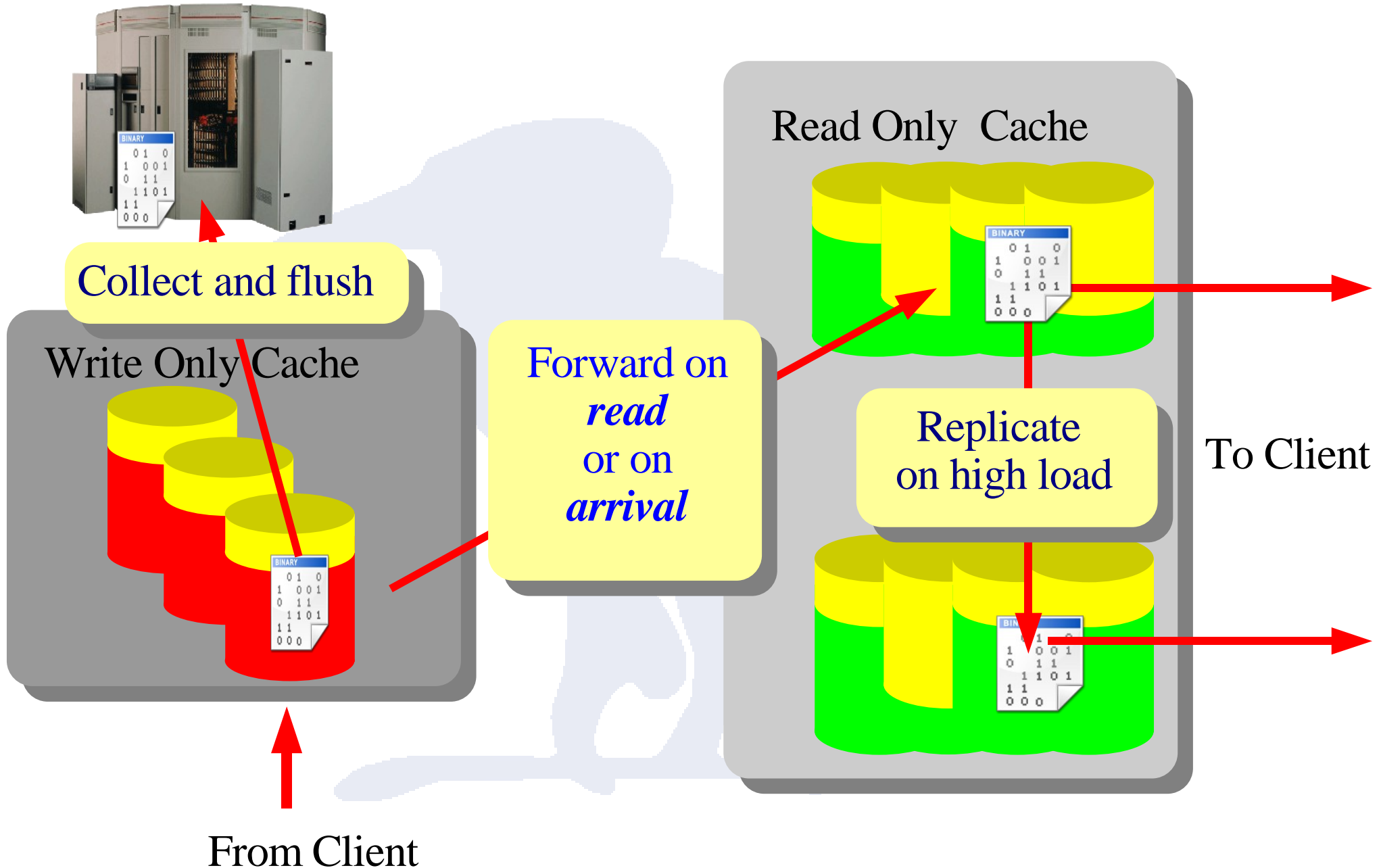




File hopping I

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

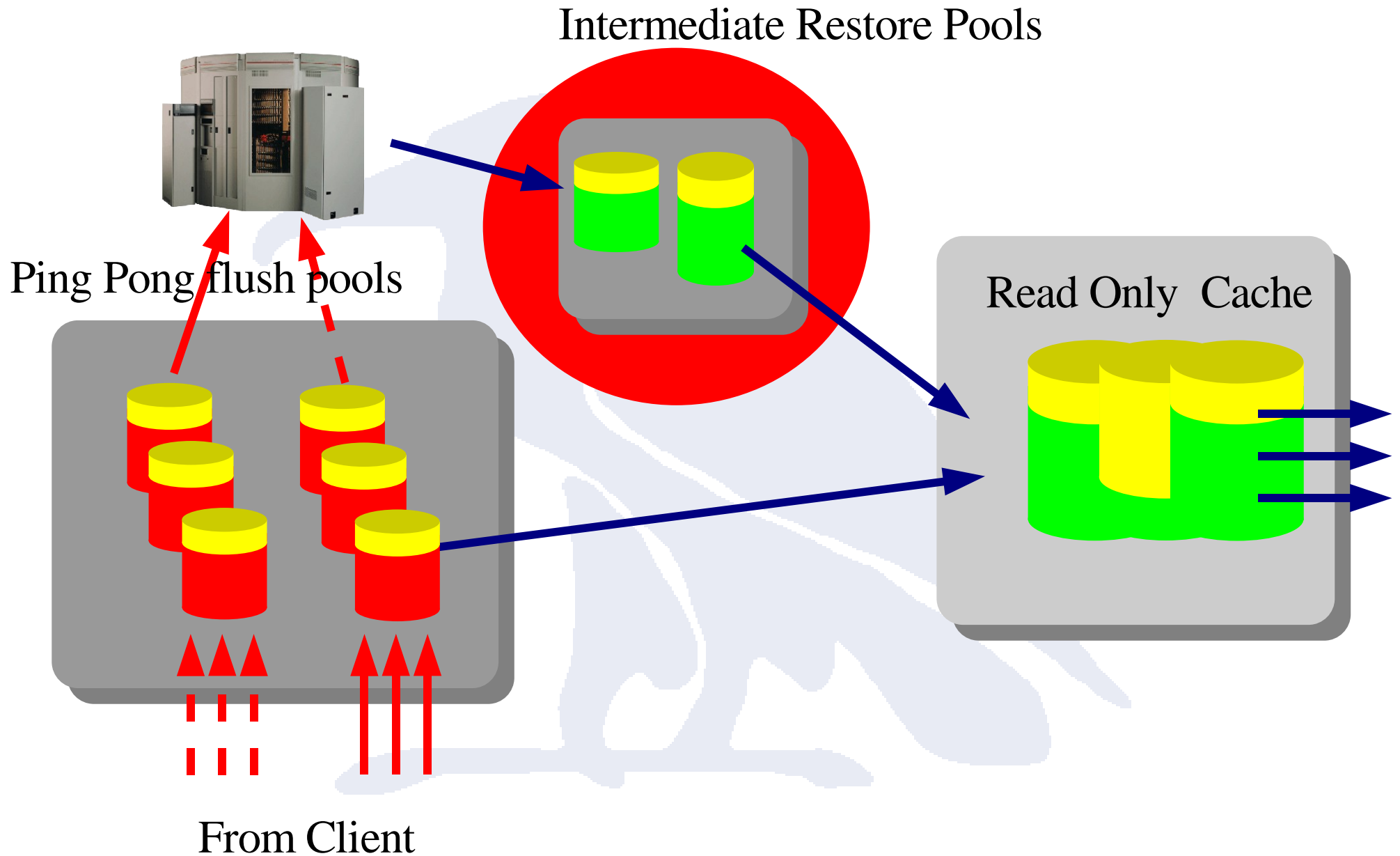


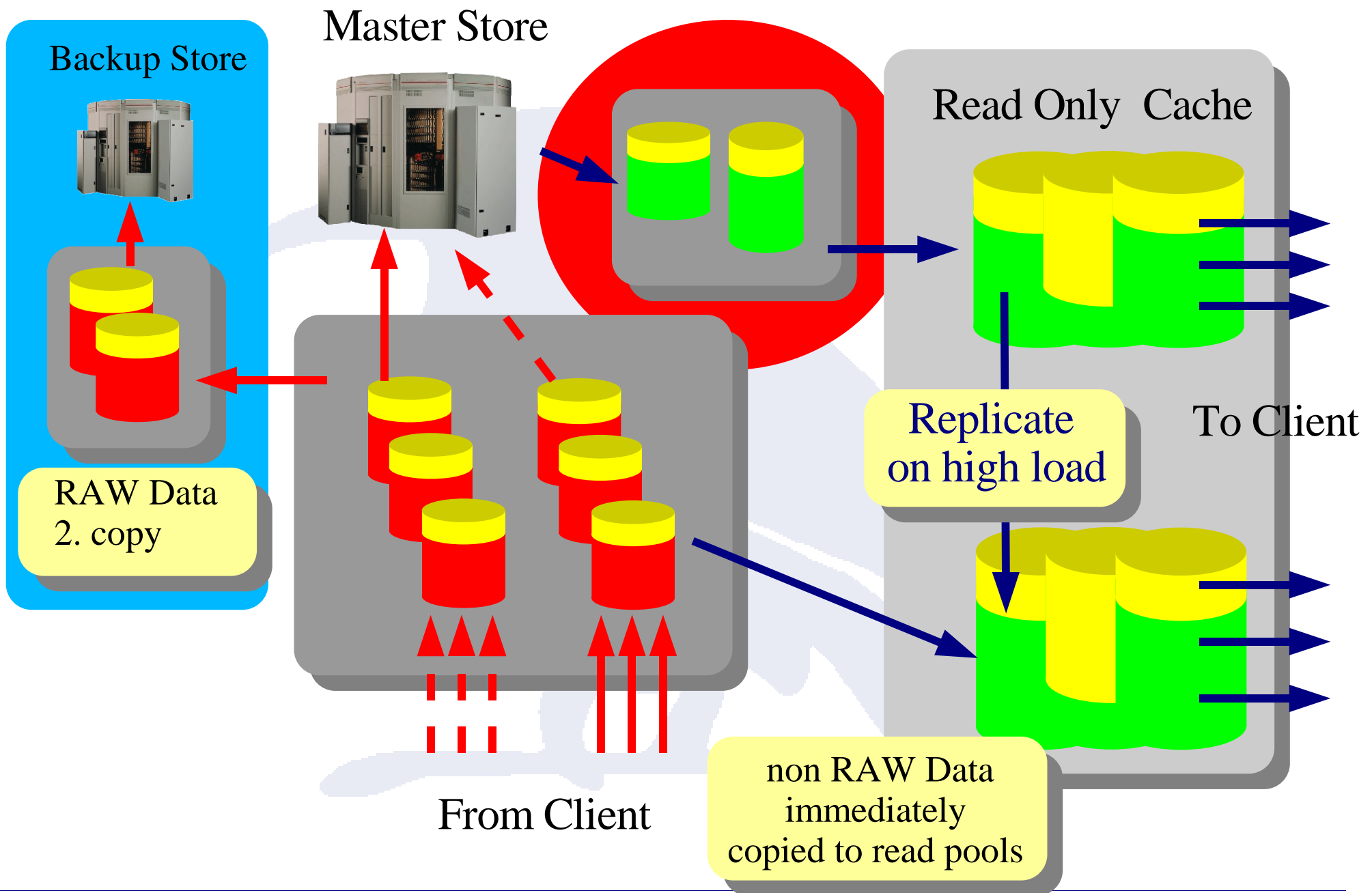




HSM interactions

- 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.





By courtesy of Alexander Kulyavtsev

Resilient dCache (pools on worker nodes)

- Controls number of copies for each dataset in dCache
- Makes sure $n < \text{copies} < m$
- Adjusts replica count on pool failures
- Adjusts replica count on scheduled pool maintenance

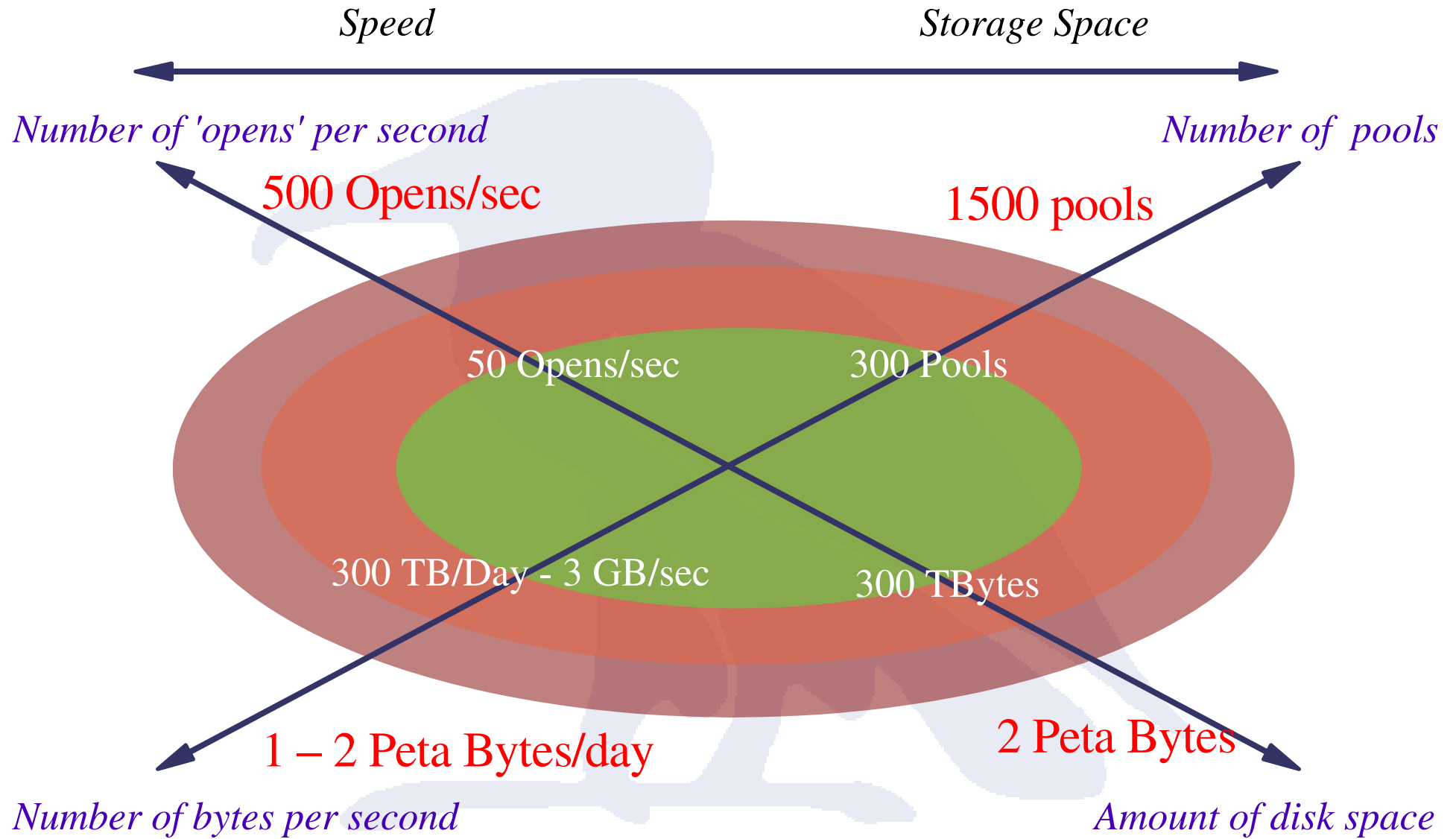
Attractive because :

- N- pool nodes may be in maintenance mode without affecting the overall availability of datasets in the dCache system.
- Improves overall performance by read striping
- Makes use of unused space on worker/farm nodes.



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 , US

IN2P3, France

SARA, NI

Nordu Grid, Sweden, Denmark, Norway ...

gridKa, Karlsruhe, Gr

RAL, UK

Triump, Canada

Tier II centers :

Germany

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

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

Italy

INFN : Bari, Torino

UK

30 % of gridPP, UK

Poland, Bulgaria, Spain

US

CMS : 7 sites, ATLAS 7 sites in preparation

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