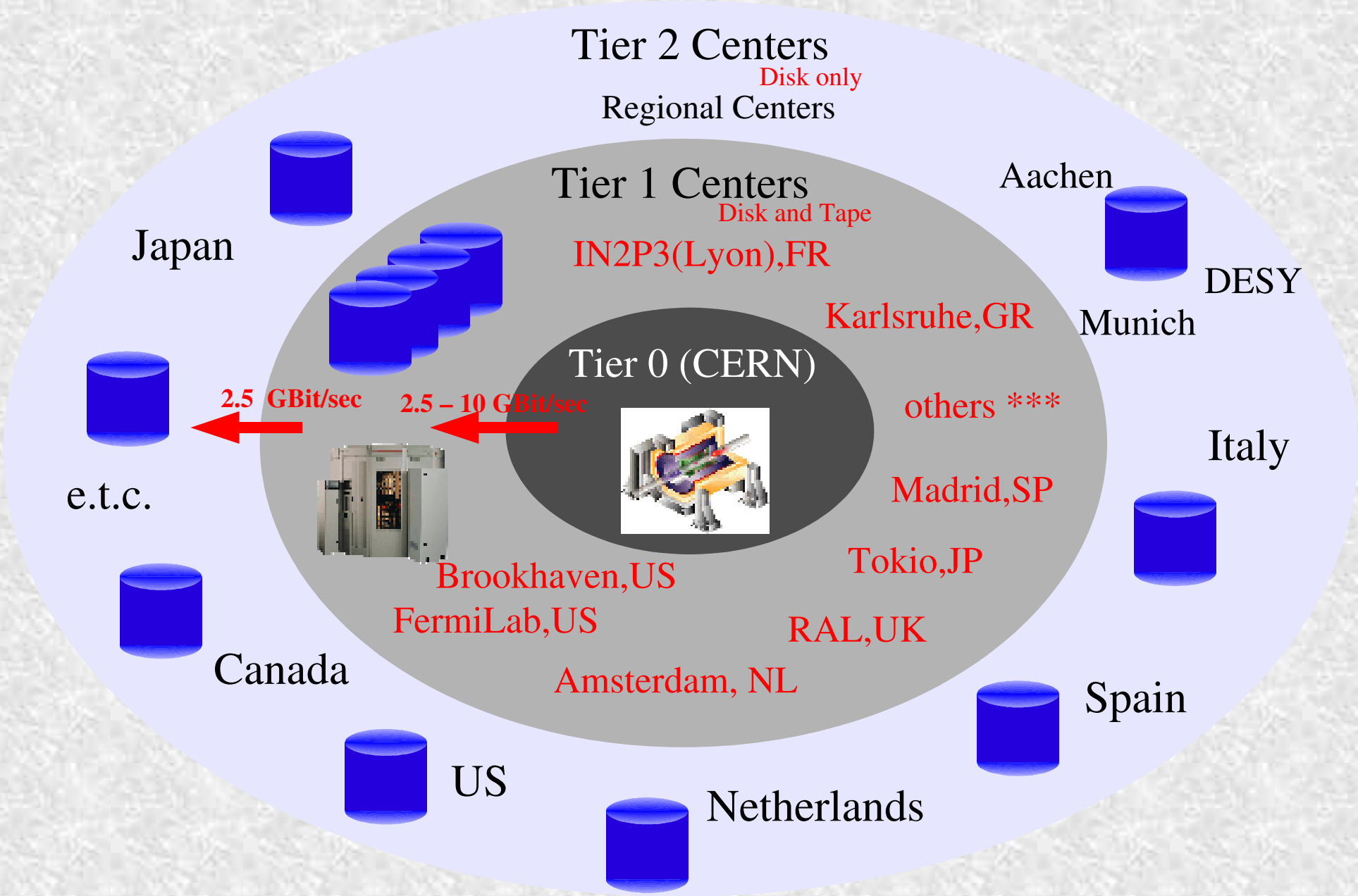
dCache, managed storage

Patrick Fuhrmann

for the dCache people


Tier 2 Centers
Disk only
Regional Centers
Tier 1 Centers
Disk and Tape
IN2P3(Lyon),FR
Karlsruhe,GR
Tier 0 (CERN)
others ***
Madrid,SP
Tokio,JP
RAL,UK
Amsterdam, NL
Brookhaven,US
FermiLab,US
2.5 GBit/sec
2.5 – 10 GBit/sec
Japan
Aachen
DESY
Munich
Italy
Spain
Netherlands
US
Canada
e.t.c.

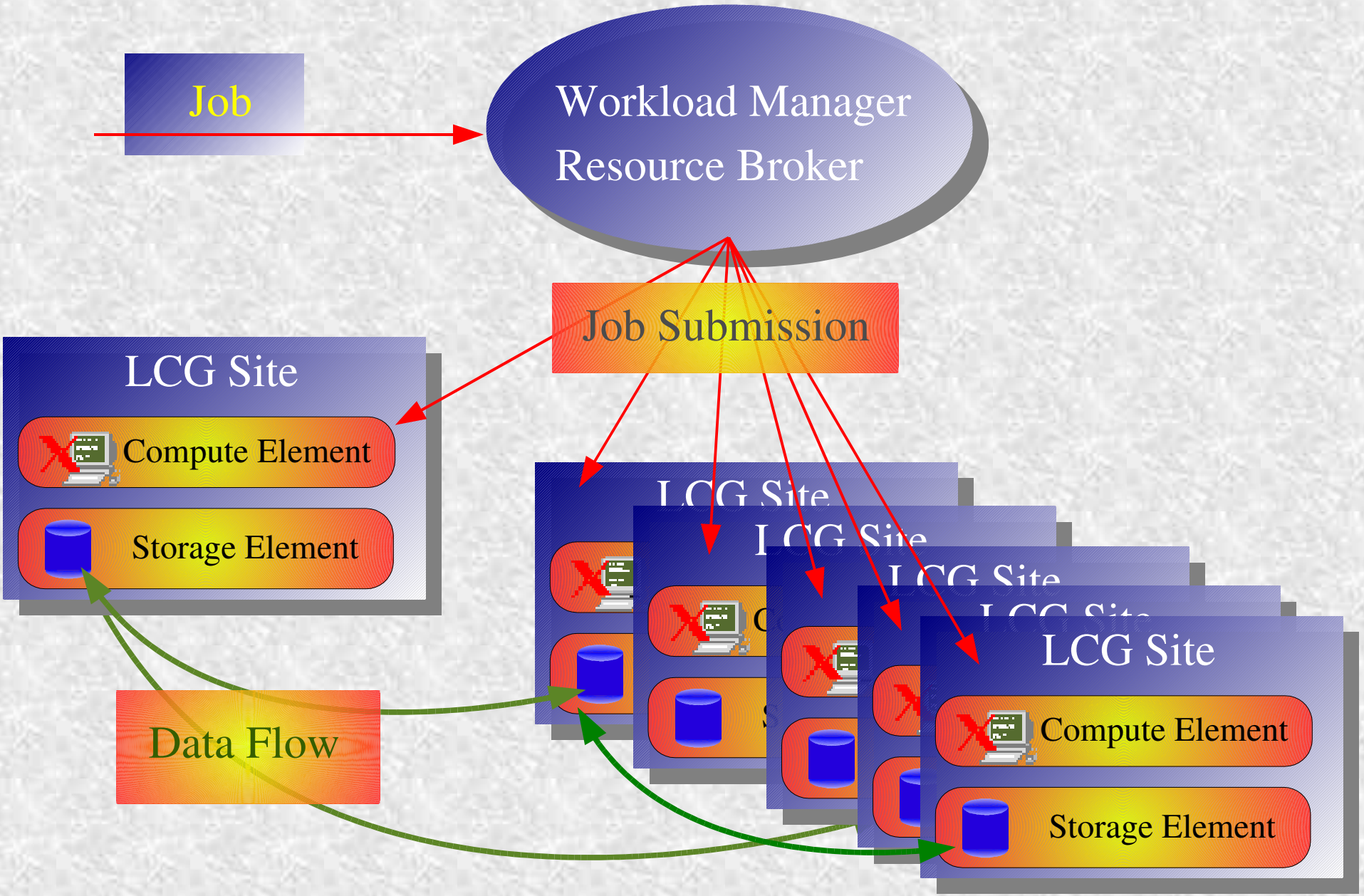
Tier I :

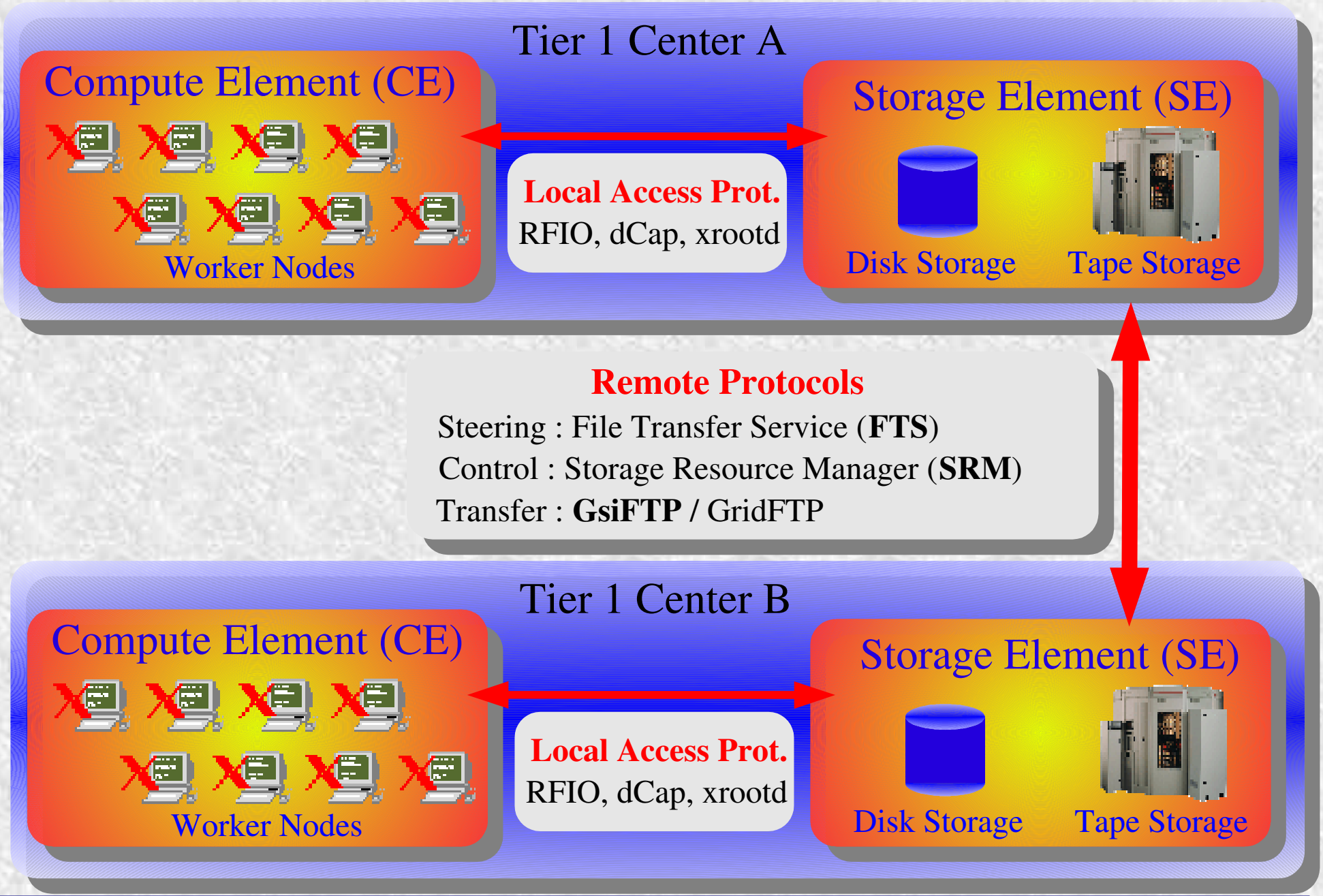
Huge systems with HSM backends

Tier II ++ :

tiny – medium size systems

w/o tape backend


Job
Workload Manager
Resource Broker
Job Submission
LCG Site
Compute Element
Storage Element
Data Flow
LCG Site
Compute Element
Storage Element


Tier 1 Center A
Compute Element (CE)
Worker Nodes
Local Access Prot.
RFIO, dCap, xrootd
Storage Element (SE)
Disk Storage
Tape Storage
Remote Protocols
Steering : File Transfer Service (FTS)
Control : Storage Resource Manager (SRM)
Transfer : GsiFTP / GridFTP
Tier 1 Center B
Compute Element (CE)
Worker Nodes
Local Access Prot.
RFIO, dCap, xrootd
Storage Element (SE)
Disk Storage
Tape Storage

Being a LCG Storage Element, means

- support of local (posix like) access protocol : **rfio, dCap, (xrootd)**
- support of the **GsiFTP** data transfer protocol

  Secure Wide Area data transfer protocol
- support of the Storage Resource Manager Protocol (**SRM**)

  Space allocation — Checksum management
  Transfer Protocol Negotiation — Dataset pinning
  Third Party Transfer on behave of dataset owner
- support of the **GRIS** information provider protocol

  Availability — Free/Available Space
  Access Protocol and access details (URIs)

The dCache Storage Element

dCache : Managed Disk Storage System

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 possible cheap disk servers.
- Supports multiple internal and external copies of a single file
- Supports 'posix like' (authenticated) access as well as various FTP dialects and the Storage Resource Manager Protocol.

Scalability

- Distributed Movers AND Access Points (Doors)
- Automatic load balancing using cost metric and inter pool transfers.
- Pool 2 Pool transfers on pool hot spot detection
- Handles bunch requests by fast pool selection unit

Configuration

- Fine grained configuration of *pool attraction scheme.*
  (write pools, subnet, directory tree, storage info)
- Pool to pool transfers on configuration of *forbidden transfers*
- Fine grained tuning : Space vs. Mover cost preference

Tertiary Storage Manager connectivity

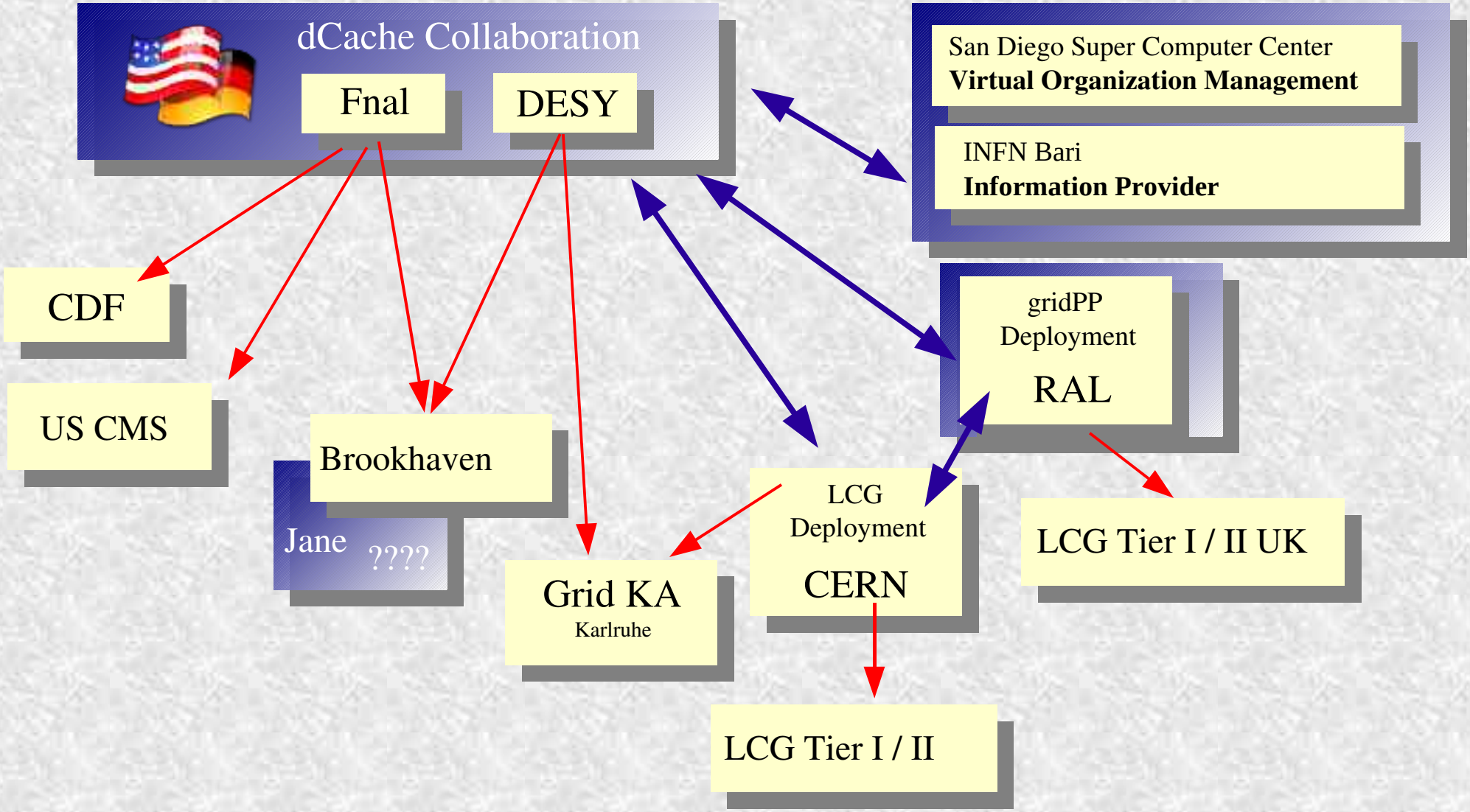
- Automatic HSM migration and restore
- HSM dCache interface by script (shell, perl ...)
- Convenient HSM connectivity for enstore, osm, TSM, Hpss

Resilient dCache

- Controls number of copies for each dCache dataset
- Makes sure n < copies < m
- Adjusts replica count on pool failures
- Adjusts replica count on scheduled pool maintenance
- Embedded farm node dCache (makes use of local disk space)

LCG Storage Element

- Local Access Protocol : **dCap**, Posix like access
- **GsiFtp** support
- **SRM** version ~ 1 (1.7) plus Space Reservation
- limited **GRIS** functionality, will improve soon


dCache Collaboration
Fnal
DESY
San Diego Super Computer Center
Virtual Organization Management
INFN Bari
Information Provider
CDF
US CMS
Brookhaven
Jane ????
gridPP
Deployment
RAL
LCG
Deployment
CERN
Grid KA
Karlruhe
LCG Tier I / II UK
LCG Tier I / II

Grid Layer

Gris

Storage Resource Mgr (SRM)

Access Layer

dCap Client

(gsi,kerberos) dCap Server

http / https

Ftp Server (gsi, kerberos)

Extended central services

Resilient Manager

HSM Flush Manager

Prestager

Core Layer

Pnfs

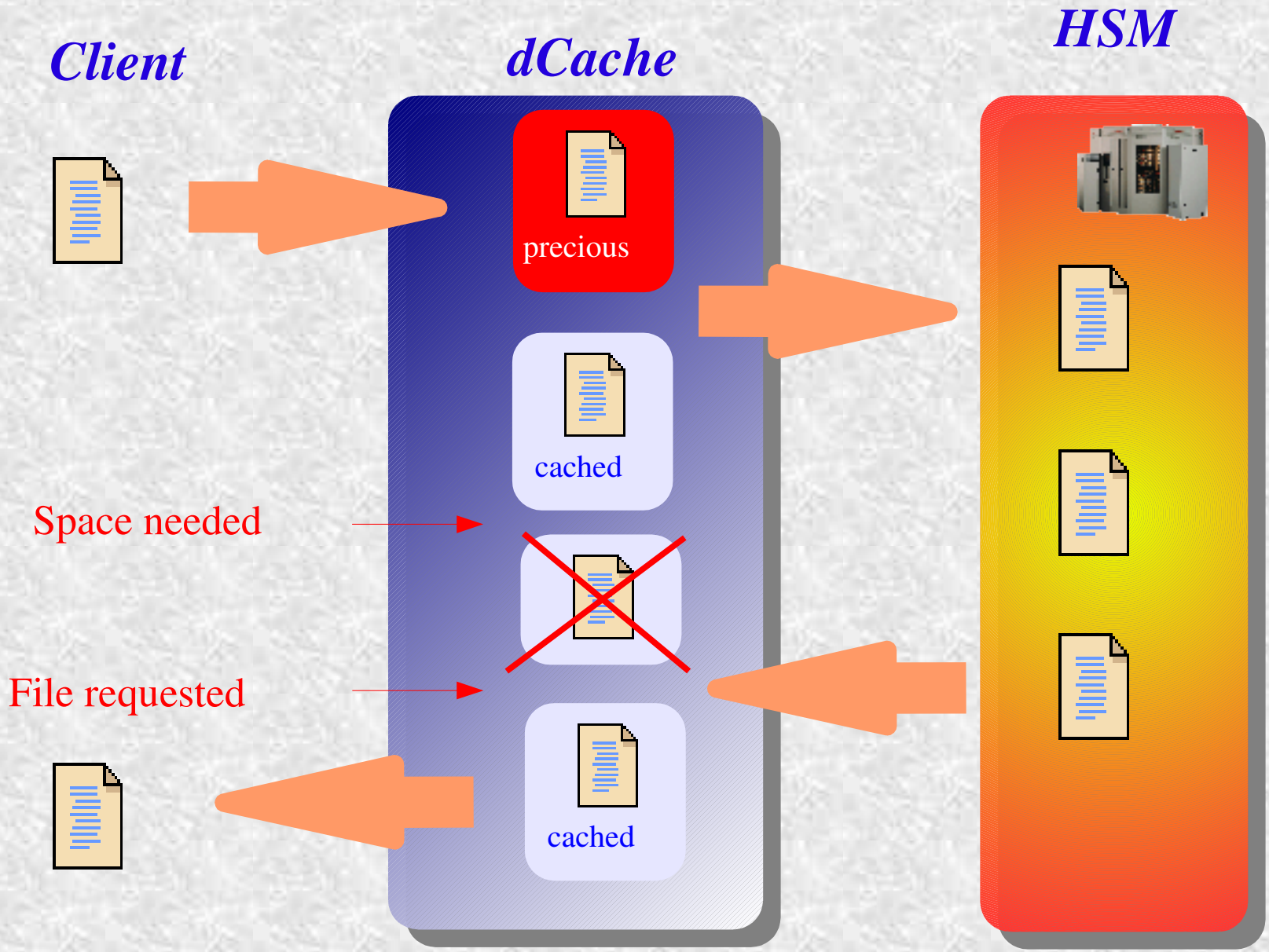
dCache Core

Cell Package

TSM Adapter

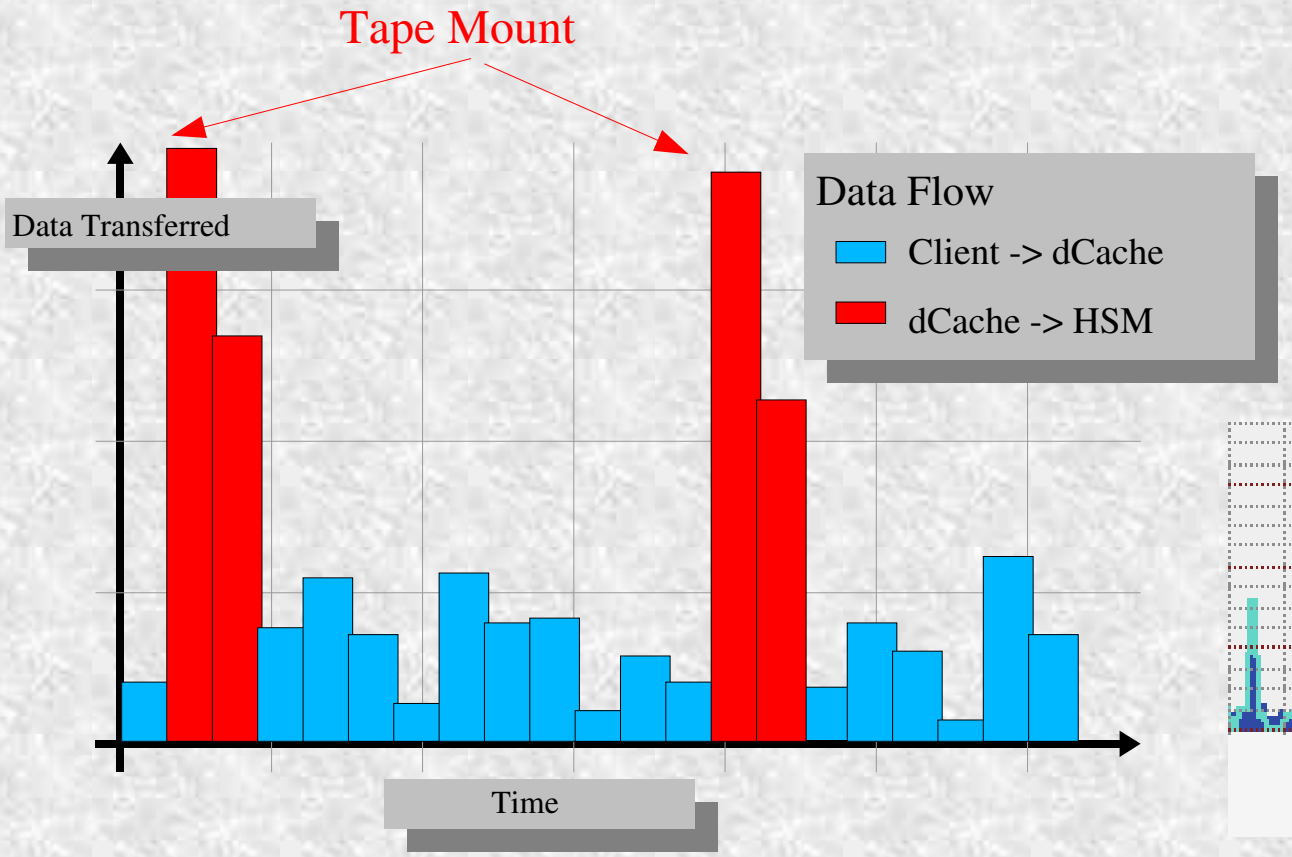
HSM Interactions

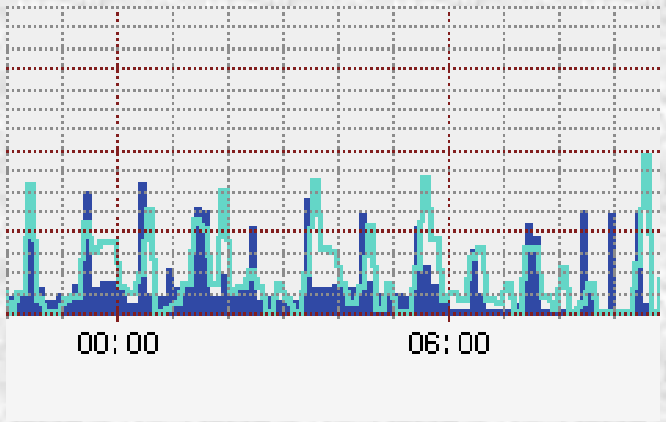
Dataset life cycle



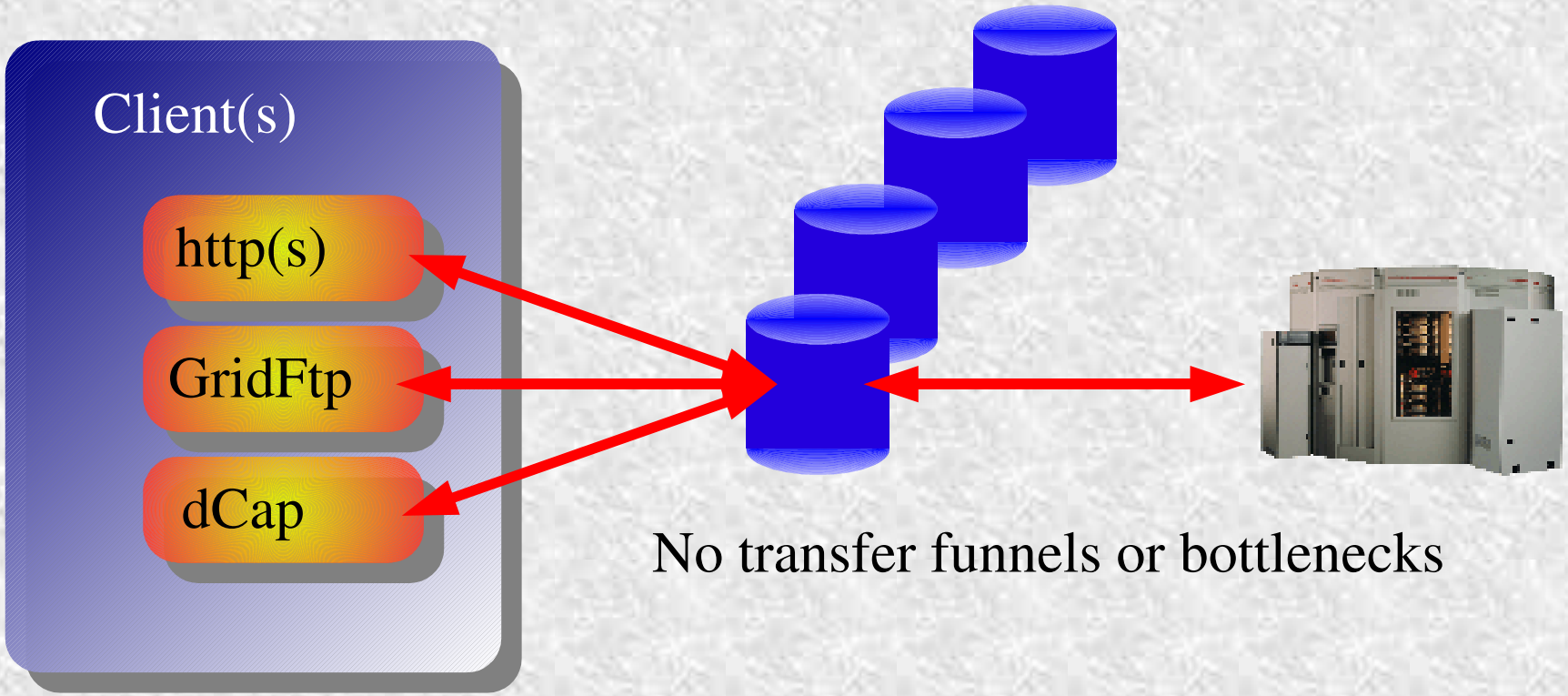
Data collected per ***Storage Class (files space, subdirectory trees)***

Collected data ***flushed*** following ***rules*** (space,time,# of files)



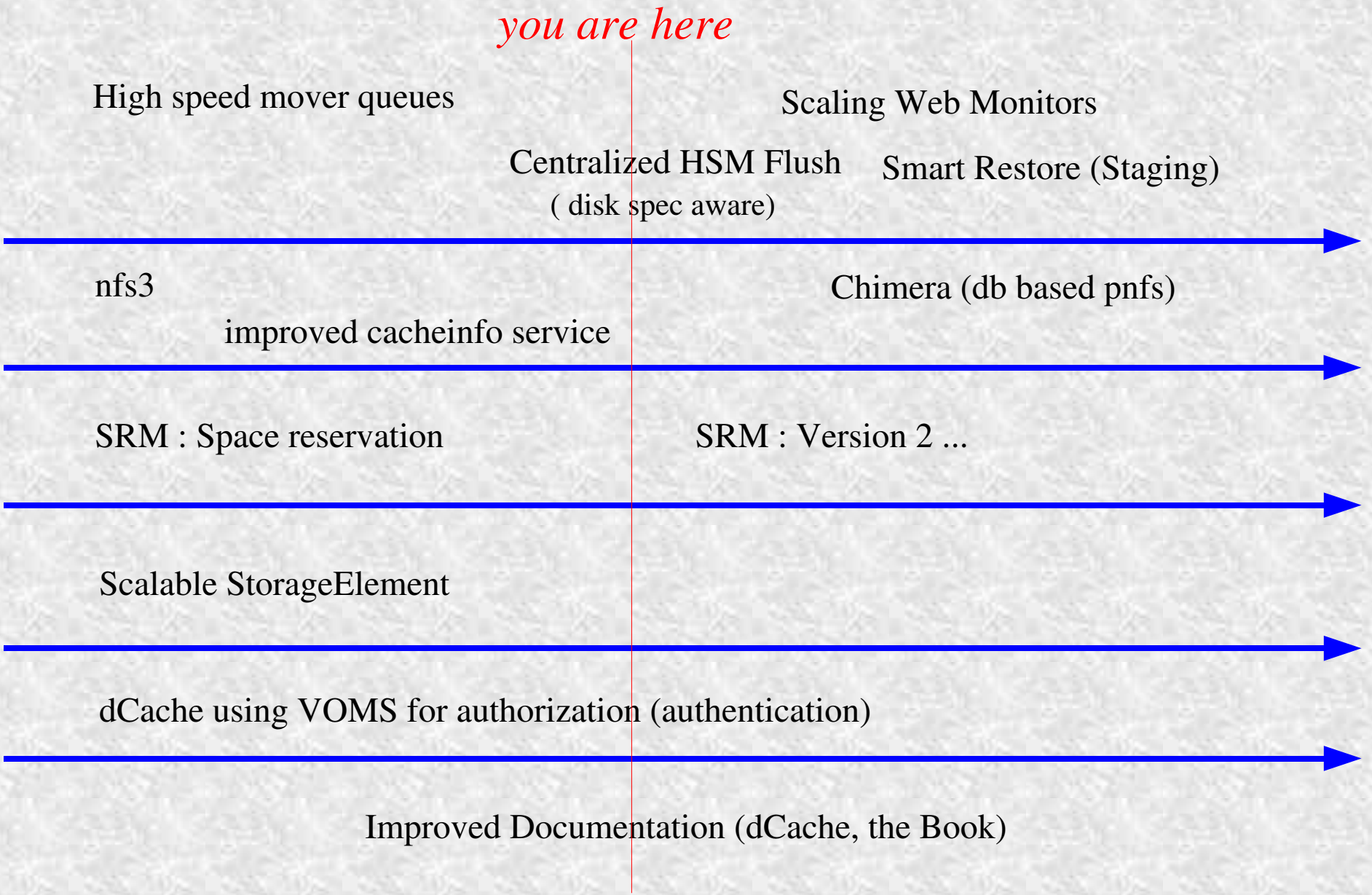


Scalability by ***highly distributed Storage Pools*** and ***direct*** client – disk and disk – hsm ***connection***



No transfer funnels or bottlenecks

Coming soon .....


you are here
High speed mover queues
Scaling Web Monitors
Centralized HSM Flush
(disk spec aware)
Smart Restore (Staging)
nfs3
Chimera (db based pnfs)
improved cacheinfo service
SRM : Space reservation
SRM : Version 2 ...
Scalable StorageElement
dCache using VOMS for authorization (authentication)
Improved Documentation (dCache, the Book)

www.dCache.ORG