dCache / Tier II workshop

Patrick Fuhrmann

for the dCache people
with contributions of

Ron Trompert, SARA
Lionel Schwarz, IN2P3
Jon Bakken, FERMI
Stijn de Weirdt
Owen Synge, gridPP
Greig Cowan, gridPP
Chris Brew, RAL Tier II
Doris Ressmann, gridKA
Reda Tafirout, Triumph
Zhenping Liu, BNL

The Team

Responsibility, dCache

Patrick Fuhrmann

Rob Kennedy

Responsibility, SRM

Timur Perelmutov

Core Team (Desy and Fermi)

Jon Bakken

Michael Ernst

Ted Hesselroth

Alex Kulyavtsev

Birgit Lewendel

Dmitri Litvintsev

Tigran Mrktchyan

Martin Radicke

Vladimir Podstavkov

Owen Synge

External

Development

Nicolo Fioretti, BARI

Abhishek Singh Rana, SDSC

Support and Help

Maarten Lithmaath, CERN

Owen Synge, RAL, gridPP

Event Flow

dCache Project Resilience

Tier II reports (gridPP, UK and GSI, Germany)

Technical News

Special use cases from your Tier I's

dCache Project Resilience

Two major development sites (DESY and FERMI)

Though each site has its development preferences,
we are using a common CVS and build system and know
each others code quite well.

Each site has at least 2 FTE with permanent contracts assigned to the project at any time to ensure continuity

Both sites have a vital interest in further support and development of dcache because both have build their own mass storage infrastructure on top of it.

dCache has becoming part of non HEP specific projects for broader acceptance and funding. (German d-grid and US SciDAC program)

SRM, The Scalable SE, Co-scheduling ...

FNAL mainly supports OSG specific issues

DESY is hosting and coordinating support@dCache.org, the user forum and the Book.

Good news : Beginning of August (lastest) DESY will get one additional FTE for user support and dCache packaging coordination.

> This will be Owen Synge from gridPP,UK. He already is dCache expert. He did the dCache YAIM integration and coordinated the UK Tier II dCache first level support. So he already is familiar with the community. (And the community with him.)

More good news: With the next dCache release there will be the Source code RPM freely available from the dCache.ORG download area. (Certain restrictions apply)

Technical News (and bits and pieces of a tutorial)

YAIM Integration

SRM WLCG V2.2 Schedule

xRoot Protocol Integration

VOMS Proxies, Roles, VO's

Resilient dCache

Multi I/O queues

Load balancing

Status of chimera

Scaling issues

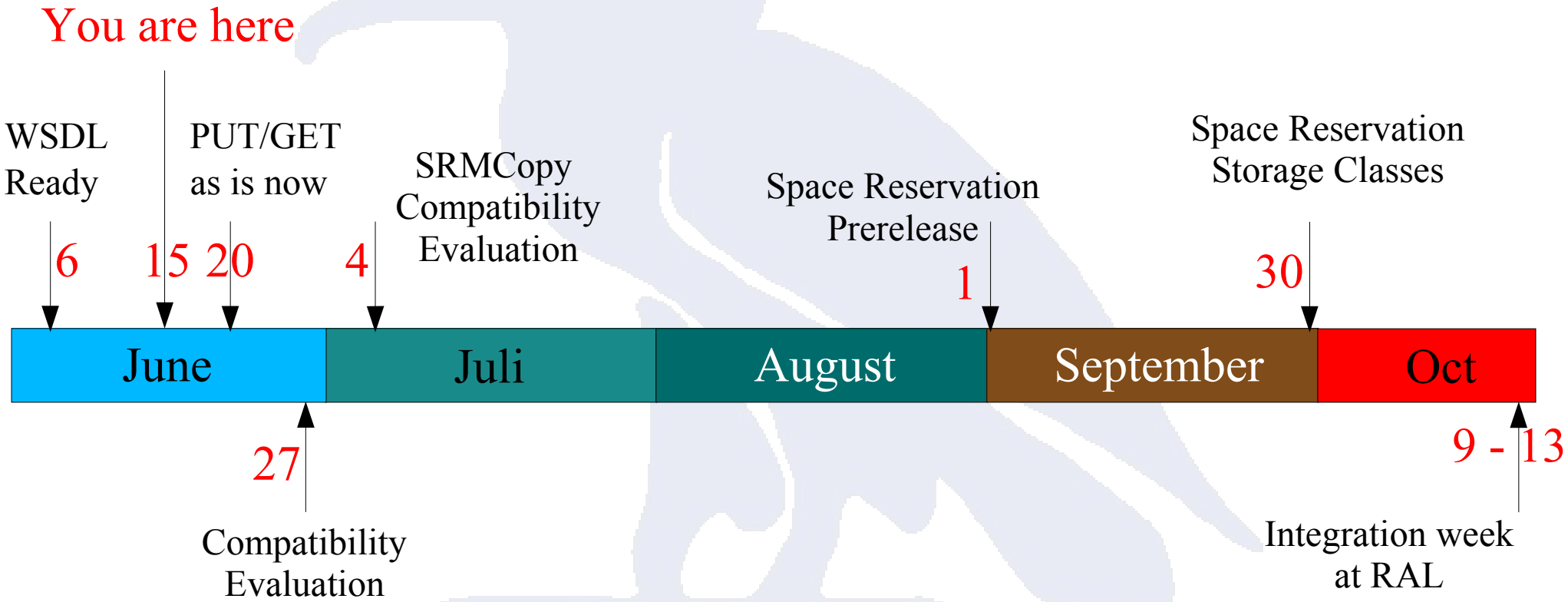
On behalf of Owen :

> YAIM is in the process of being integrated into the dCache software itself. The YAIM scripts, in the future, only need to point to our scripts.
>
> YAIM supports the information system in all gLite releases.
>
> Together with gridPP, we are building an 'automated functional regression test system' which will reduce the times between bugfix releases.

Patricks remark :

The bottom-line is certainly that we are going to put more effort into dCache deployment, installation and YAIM integration.

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



Basic xRoot Protocol done

- xRoot Protocol Door and Mover integrated (dCache Native)
- Takes full advantage of dCache cost mechanism.
- Pools can be configured to have special xRoot queue.
- xRoot requests may be directed to special xRoot pools.

In Progress

- No Authentication yet (but access can be set 'read-only')
- Discussion with xRoot security experts ongoing.
- OLBD protocol implemention still in discussion.

- Beta release available as patch to dCache 1.6.6-5 on www.dCache.org
- Tested by GSI and at 2 gridPP sites (thanks to Kilian, Victor and Chris Brew)
- more evaluation sites are welcome

VOMS Extended Proxies supported by dCache through gPlazma

gPLAZMA (grid-aware PLuggable AuthoriZation MAnagement)

gsiFtp and SRM already gPlazma aware

gsiDCap in preparation

LCMAPS files not yet supported (done: dcache.kpwd file and GUMS,SAS)

By courtesy of Alexander Kulyavtsev

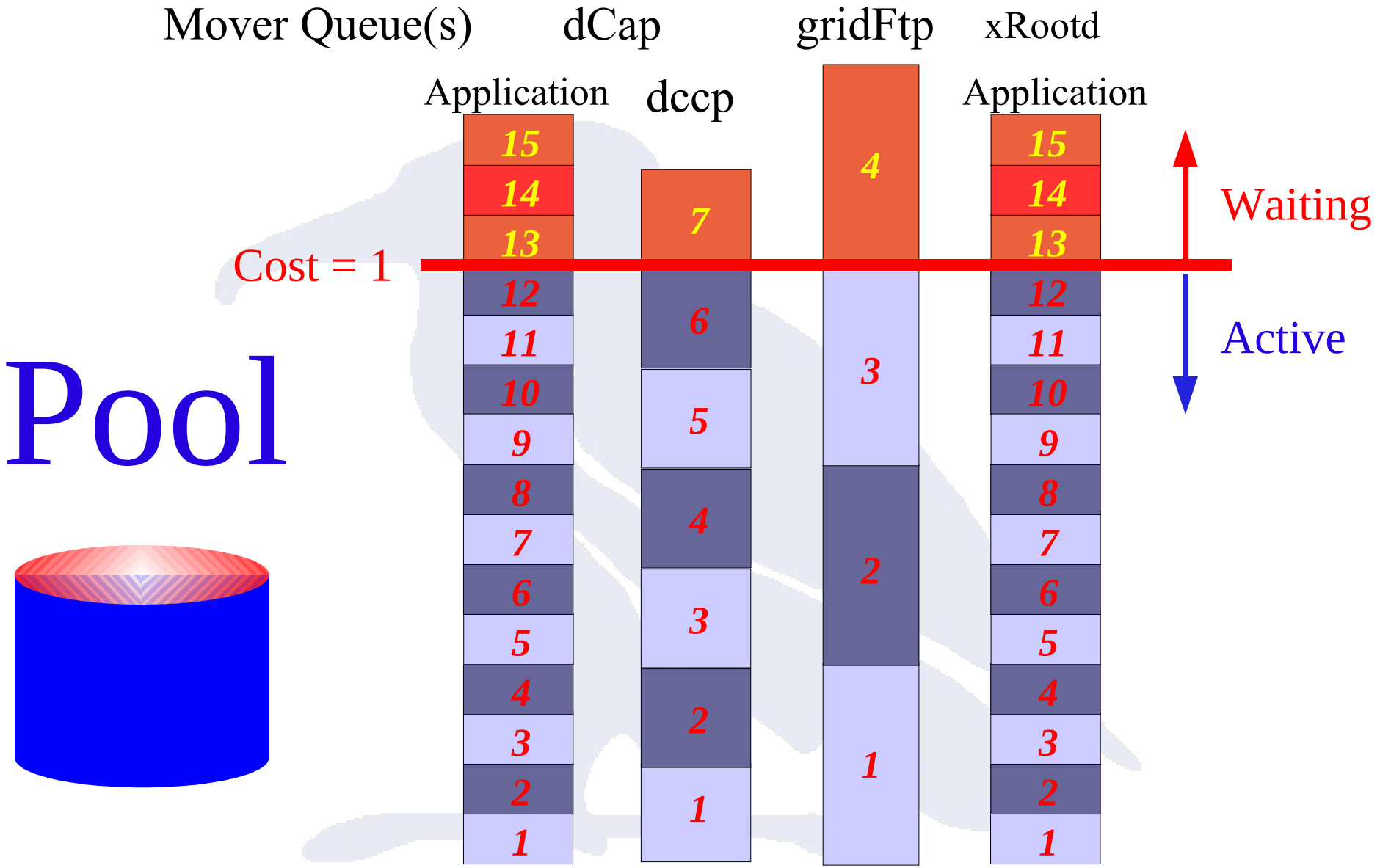
Resilient dCache

- 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
- *you may run resilient and classic dCache within one dCache instance.*

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

The resilient dCache module is undergoing a major redesign phase in order to fix known problems and meet enhanced requirements. The new code will not be part of 1.7.0. Please contact support@dCache.org if you intend to install a resilient dCache within the near future.



1.7.0 should have 2 (fast and slow) queue in standard setup.

Load balancing may be tuned depending on admin's preferences.

- You want to fill up empty space first rather than have the same number of movers running on all pool nodes :

  set pool decision **-cpucostfactor=0.0 -spacecostfactor=1.0**

- SC : You may want to have nice mover balancing but don't care to much on equal space distribution :

  set pool decision **-cpucostfactor=1.0 -spacecostfactor=0.0**

- You prefer random selection of pools (will be in 1.7.0)

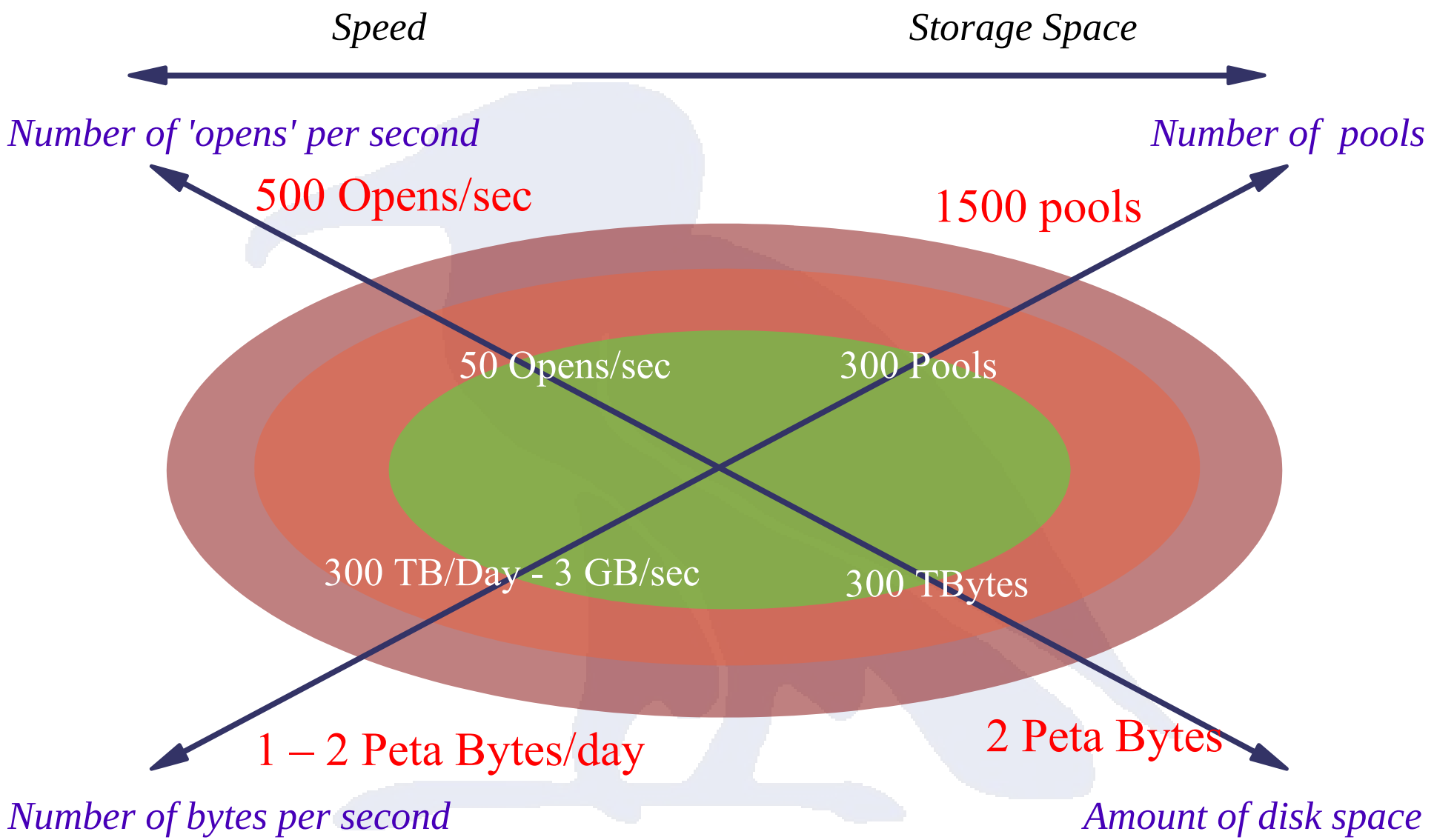
  set pool decision **-cpucostfactor=0.0 -spacecostfactor=0.0**

With 1.7.0 these values may be set per pool group. (see dCache partitioning)

The overall load is balanced by dCache automatically using p2p transfers.

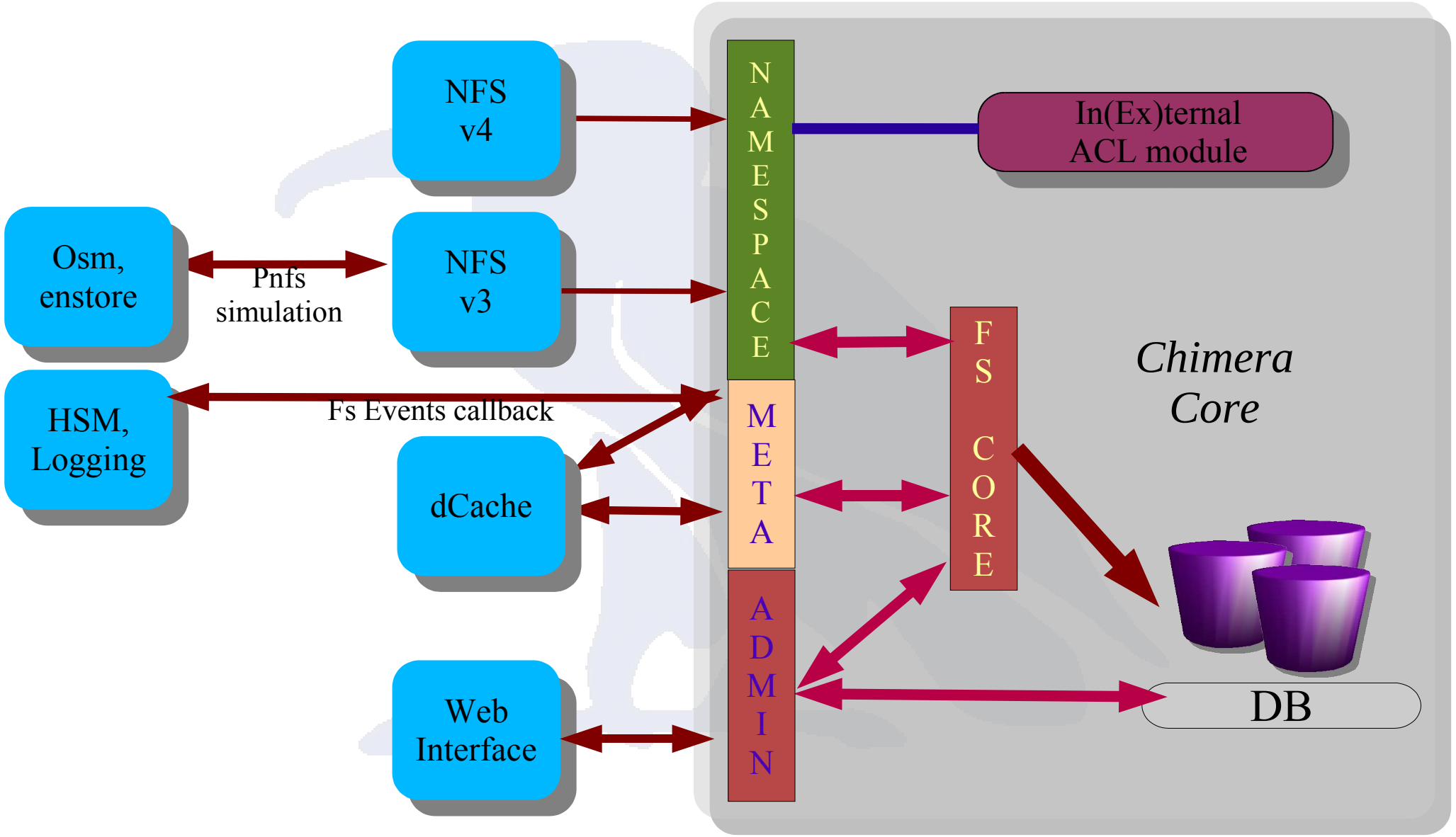
- you determine the threshold on which p2p starts on file access hot spot detection.
- starting 1.7.0 this is part of the dCache partitioning schema
- (HSM only) you may specify a threshold from which on p2p is stopped to protect the system and instead fetches file from HSM.

Scaling issues



Chimera Layout

by courtesy of Tigran Mkrtchyan



Coding finished

First design review together with FERMI : done

Larger evaluation is in progress

First production installation selected (they don't know yet)

Looking for evaluation sites to help us (non production only)

VO , user quotas

No promises here, some logical problems not solved on ownership of replicas

Current Recommendation : .5 – 1T Pools assigned to VO's

Data Movement Admin Tools (CopyManager, draining pools)

Improved version will come with 1.7.0
(File destination will follow PoolManager rules)

SRM 2.x prereleases

We'll try to make this available to selected sites, though the final version might differ significantly

DN not recorded for srm transfers when your dCache is destination

FIXED

dCache GridFTP logs can't be easily published into R-GMA

Long standing issue. We have all the required infos at one location in dCache available now. Still the r-gms interface is missing. Help would be appreciated.

day to day maintenance tools

Long standing issues. We have all the required infos available now. Still the r-gms interface is missing

- Pools may now be selected by **Protocol** in addition to IP address, file system location and dataflow direction.
- dCache can be partitioned, having different sets of parameter (and with that behaviour) on different sets of pools. (The Book)
- See Book for advice how to migrate from classic SE to dCache.

HSM Specific

- HSM decouple pool sets
- Smart flushing (stream control)

Pnfs(postgres) database backup

Run slony : you will have continues backup

or Postgres incremental backup

Supported Platform :

pnfs (c-code) : solaris, linux, aix, irix (Darwin)

Chimera (new pnfs) will be java anyway

Rest of dCache : find a jvm

Some Tier I details

Detailed report by Jon Bakken, Lionel Schwarz and Ron Trompert
available at www.dCache.ORG

FERMI (CMS Tier I setup)

Fermilab

By courtesy of Jon Bakken, CMS, Fermi

Head Nodes

PoolManager
- Admin Door
- LocationManager
- hsm Controller

Pnfs
- Pnfs Server
- Pnfs Manager
- Pnfs Postgres DB's

Resiliency
- Replica Manager
- Replica Postgres DB's
- gridFtpDoor

More Doors

Management
- 6 * CopyManager
- 5 * dCap Door
- 1 * gridFtpDoor

Control Door(s)
- SRM - v1
- SRM - v2
- PinManager
- SpaceManager

Information System
- Billing (DB) + Postgres
- HTTPS (deamon 2288)
- Info Provider

HSM Decouple Space
8 Nodes
9 TBytes

Read/Write Space
23 Nodes ; 100 TBytes
New this Summer :
100 Nodes ; 600 TBytes

Worker Node (Resilient)
500 Nodes ; 55 TBytes

Typical Pool Node (non resilient)

By courtesy of Jon Bakken, CMS, Fermi

1 * gsiFtp door

AND 1 * Infortrend with 2 partitions

OR 1/2 * nexsan with 4 partitions

each partition runs 2 pools

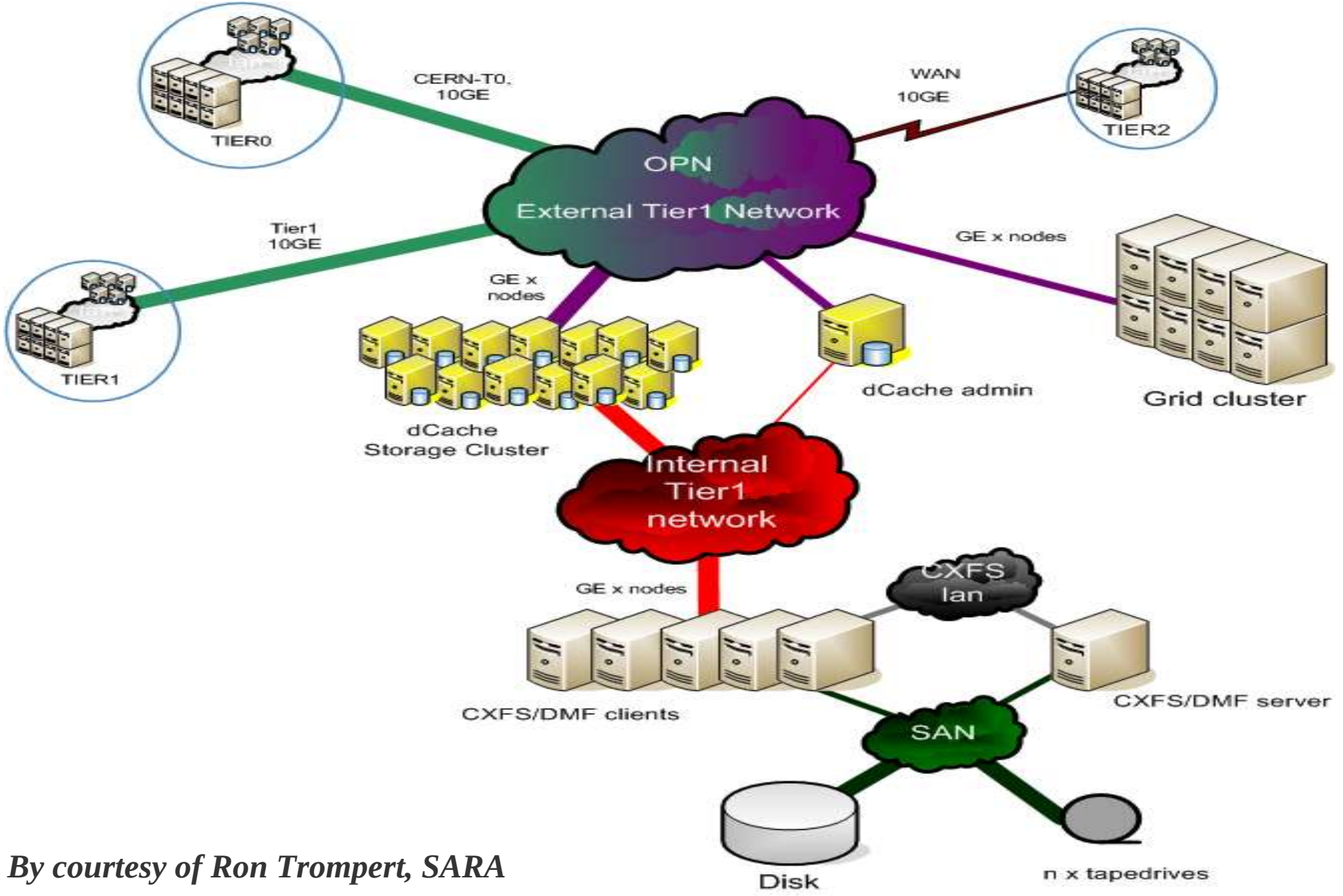
- 10 Gbytes for volatile data
- rest for production data

each pool runs 2 mover queues

- LAN : max 600 active movers per pool (local dCap random I/O)
- WAN : max 3 active movers per pool (gsiFtp streaming)

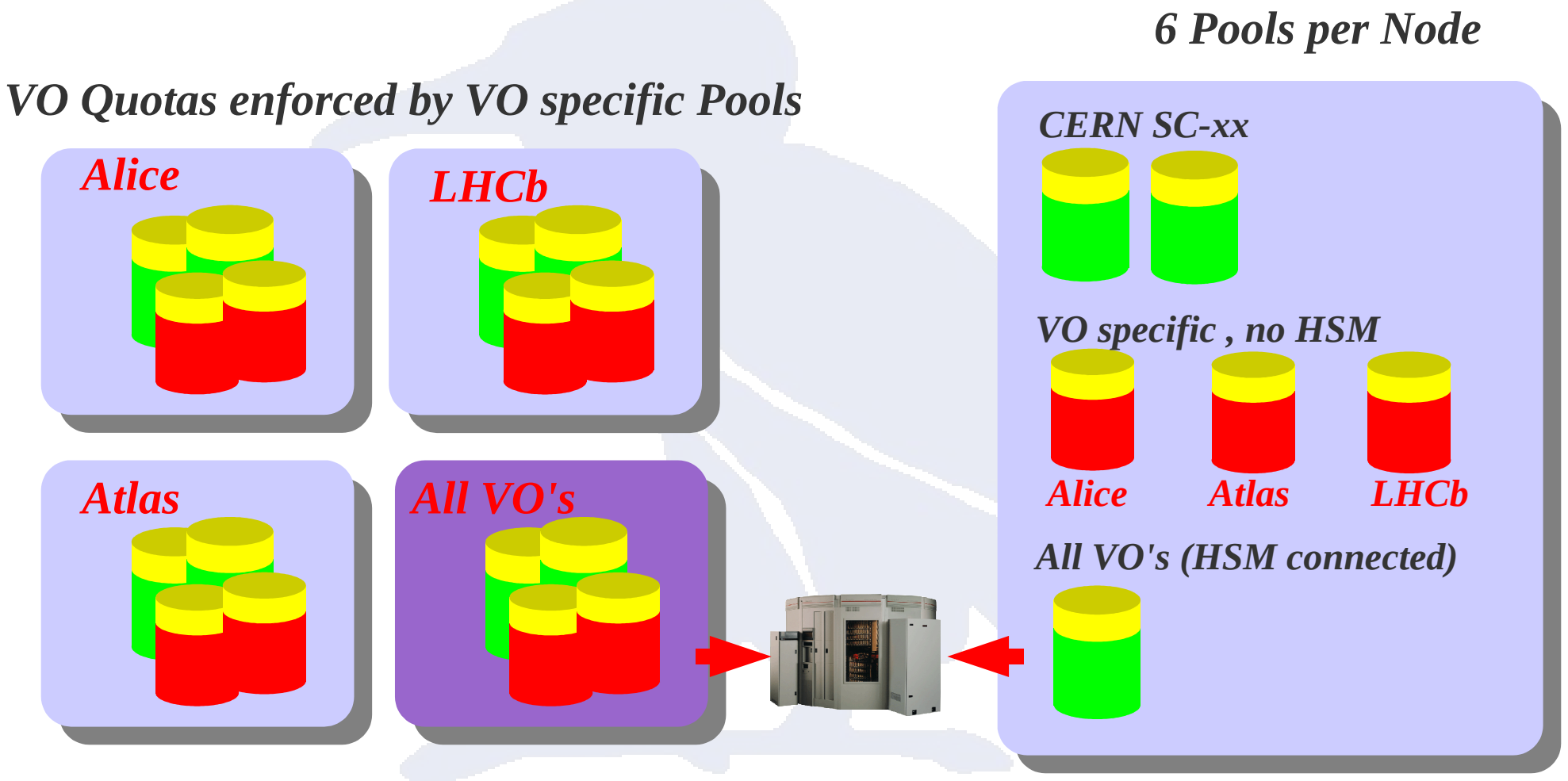
More details :

- Each node 2 bonded GE interfaces
- CPU : dual Opteron with dual core



By courtesy of Ron Trompert, SARA

Node / Pool configuration => Quotas :-)



Tier I summary

	Headnodes	Movers		Per Node		
		gsiFtp	dCap	Pools	NetIf	Space/TB
FERMI	7	3	600	7	2 (Bonded)	4 – 6
SARA	1 -> 2	6 (1 queue only)		7	1	
IN2P3	1 -> 2	15 - 20	300	7	2 (HSM/Ext)	~ 3
gridKa	1 + SRM	5 - 50	300	2	2(Intern/Extern)	1 - 1.5
BNL	1 + PNFS	8–10 (1 queue yet)			1(open firewall)	0.100 - 1
Triump	1 + PNFS	15 (1 queue yet)		3	1(no firewall)	0.2 - 0.8
NorduGrid	very special setup					

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