

DESY DV-Seminar

16. Juni 2008

Hamburg



The dCache Storage Element and it's role in the LHC era

Martin Radicke
for the dCache team



Topics for today

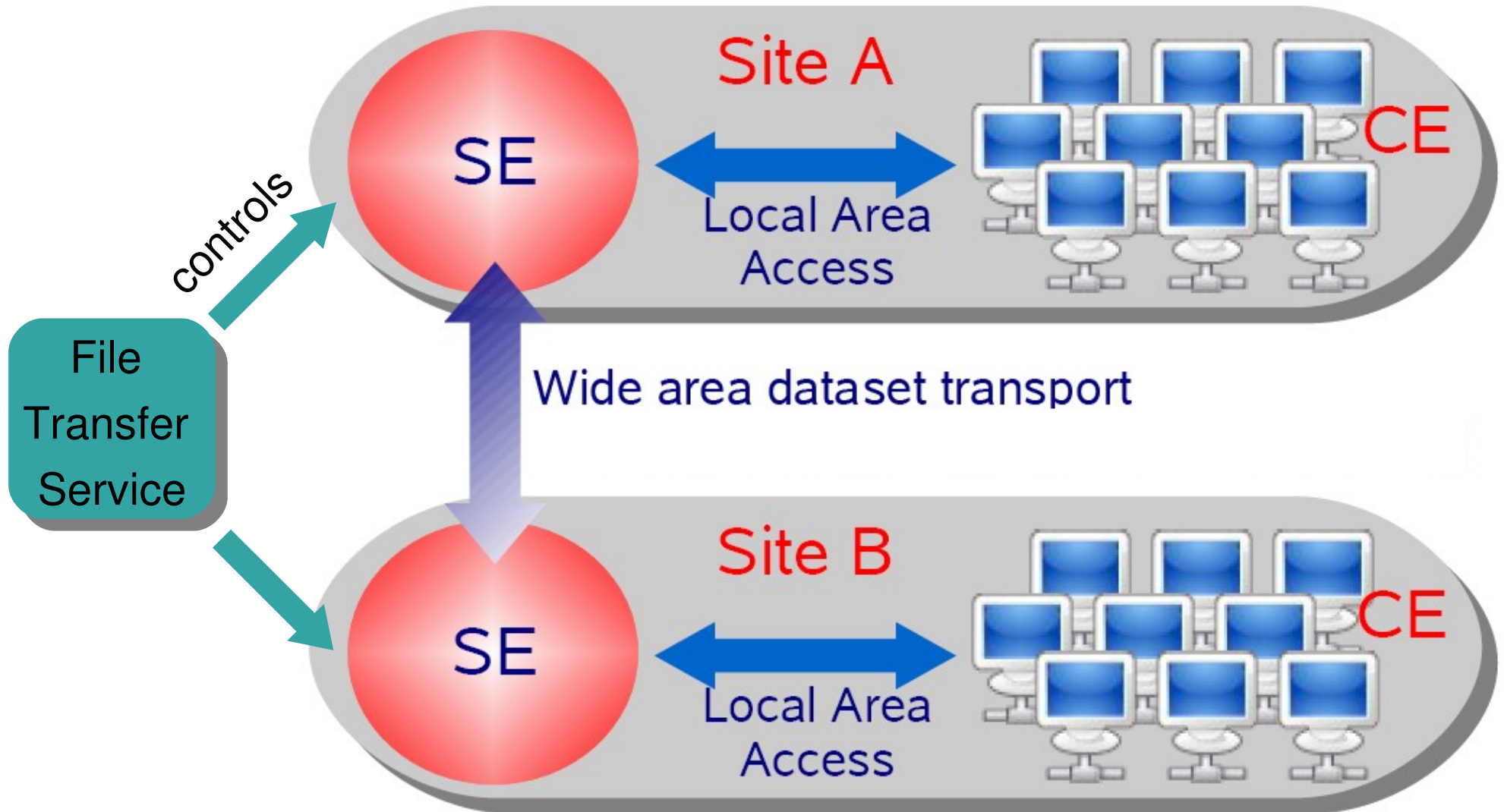
- ▶ Storage elements (SEs) in the grid
- ▶ Introduction to the dCache SE
- ▶ Usage of dCache in LCG and @ DESY
- ▶ Project layout
- ▶ current and future developments



Storage Elements in the grid



Introducing the Grid StorageElement (SE)





Requirements for a Grid SE

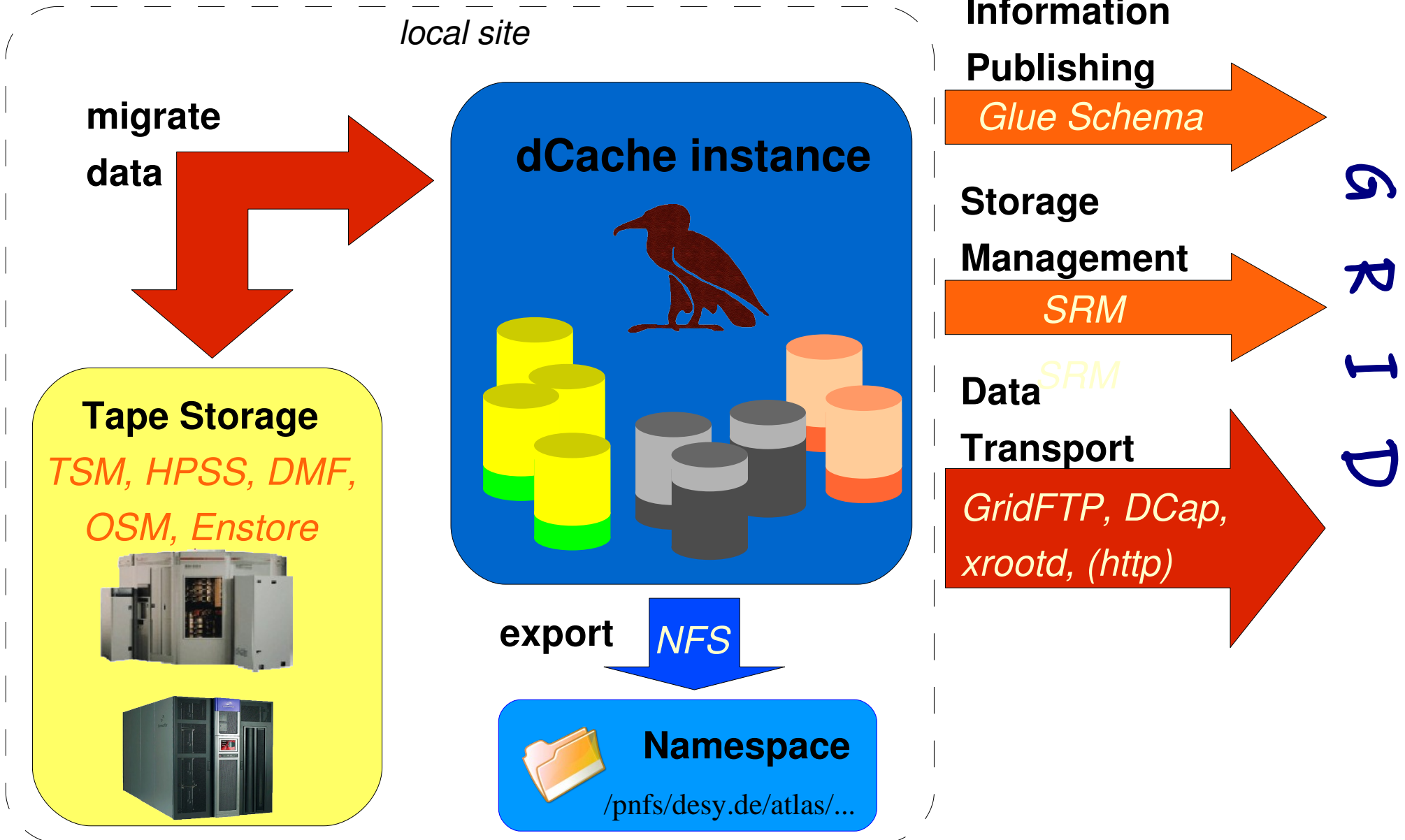
- ▶ serving large amounts of data locally
 - Posix-like random access from worker node
 - huge number of parallel file requests, low latency required
- ▶ exchanging big datasets across datacenters
 - file streaming from/to remote SEs (Tier centers)
- ▶ managing storage
 - space reservation for guaranteed streaming bandwidth
 - space attributes (tape only, tape + disk, disk only)
 - transport protocol negotiation
- ▶ information publishing
 - available space (max, reserved, used), type (disk, tape)
 - Which VO owns which space?
 - Service endpoints (URLs)



The dCache SE



The dCache SE – a bird's view





dCache key concepts

- ▶ managed storage on off-the-shelf hardware
- ▶ combines hundreds of commodity disk servers to get a huge PetaByte-scale data store
- ▶ single-rooted namespace to give users a FileSystem-like view to the storage
- ▶ strictly separates between namespace and data repositories → increased fault tolerance
- ▶ allows several copies of a single file for distributed data access
- ▶ internal load balancing using cost metrics and inter-pool-transfers



▶ Load balancing and overload protection

- IO-Request Scheduler
- storage pool selected by Client IP, Protocol, Directory and in a 2nd round by load and free disk space
- IO-queues on pool (per protocol)

▶ File hopping on

- automated hotspot detection
- configuration (read only, write only, stage only pools)
- on arrival (configurable)
- outside/inside firewalls

▶ Replica Management

- system ensures:
 $n < \text{file copies} < m$



▶ local, posix-like access

- DCap – dCache's native protocol
 - client (-library) provided by dCache
 - GSI authentication supported
- xrootd - transparent access from within ROOT
 - GSI authentication underway

▶ remote, streaming access

- GsiFTP – the Grid standard
 - active, passive
 - multiple streams
- HTTP(s) – under discussion
 - read-only prototype available



Storage Resource Manager (SRM)

- ▶ an standardized interface to access grid SEs
- ▶ prepares file transfers
 - client ↔ SE (file up- and download)
 - SE ↔ SE (3rd -party transfer across grid sites)
 - negotiates the transfer protocol (mostly GsiFTP)
- ▶ space management
 - dynamic or static space reservation for guaranteed streaming bandwidth during data taking
 - spaces have attributes attached
 - Access Latency and Retention Policy steering physical file location
- ▶ version 1 and 2.2 supported by dCache



Access to Tertiary Storage

- ▶ dCache has a simple yet powerful API to connect to Tape Backends
 - hooks for file write, file read and file delete
 - currently used with: TSM[®], HPSS[®], DMF[®], OSM, Enstore
- ▶ coordinated migration to tape (“flushing”)
 - collect files and flush in bunches, based on time, total size and/or storage class
- ▶ currently each pool autonomous
- ▶ under development: central Flush-and Stage-Manager
 - better disk-and tape -utilization





Project layout



The Project: People and Funding

Head of dCache.ORG

Patrick Fuhrmann

Head of Development FNAL :

Timur Perelmutov

Head of Development DESY :

Tigran Mkrtchyan

Core Team (Desy and Fermi)

Andrew Baranovski

Bjoern Boettscher

Ted Hesselroth

Alex Kulyavtsev



Iryna Koslova

Dmitri Litvintsev

David Melkumyan

Dirk Pleiter

Martin Radicke

Owen Syngé

Neha Sharma

Vladimir Podstavkov

External

Development

Gerd Behrmann, NDGF

Jonathan Schaeffer, IN2P3

Support and Help

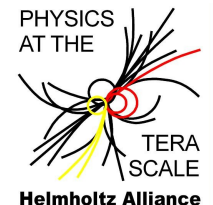
Abhishek Singh Rana, SDSC

Greig Cowan, gridPP

Stijn De Weirdt (Quattor)

Maarten Lithmaath, CERN

Flavia Donno, CERN





- ▶ dCache development infrastructure @DESY
 - SVN source code repository
 - continuous integration system
 - automated building and (distributed) testing
 - testbed for integration into gLite distribution
 - weekly video conferences with FermiLab and NDGF
- ▶ deployment
 - binaries directly from www.dcache.org or through YUM repositories
 - a very stable version as part of Cern's gLite
 - source code available under special open source license



▶ documentation

- book/wiki/publications: www.dcache.org

▶ general support

- trouble ticket system: support@dcache.org
- user forum: user-forum@dcache.org




▶ German/International support group (as part of HGF Alliance) resolving GGUS tickets

▶ DESY provides direct phone support

- weekly with all Tier-1 sites
- adhoc with Tier-1 centers in Brookhaven (BNL), Karlsruhe (FZK) and Lyon (IN2P3)



dCache is part of the D-Grid initiative

- ▶ development FTE through 
- ▶ mainly integration/support through DGI (1+2)
- ▶ some installations in AstroGrid and C3-Grid 
- ▶ sustainable documentation in coop. with 
 - multimedia talks
 - e-learning modules (e.g. “configuration of SRM 2.2”)
- ▶ hands-on workshops
 - installation, configuration and administration training
- ▶ preinstalled, preconfigured dCache as a virtual machine image for the desktop
 - for testing/training purposes
 - contains full gLite UI for transfer tests +VOMS support



Who uses dCache?



dCache and the LHC grid

▶ dCache used in **8 Tier-1** centers

- FZK (Karlsruhe, GR)
- IN2P3 (Lyon, FR)
- BNL (New York, US)
- FERMILab (Chicago, US)
- SARA (Amsterdam, NL)
- PIC (Spain)
- Triumf (Canada)
- NDGF (NordGrid)



▶ .. and about **60 Tier-2s**

▶ currently largest sites (BNL, FermiLab)

- 2.5 PB on disk, 2 PB on tape
- expected to scale up to > 20 PB around 2011

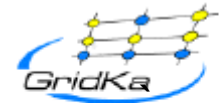
dCache will hold the largest share of LHC data



dCache @ the German Tier-1

Last hour (2008/06/15 11:00 - 2008/06/15 12:00)

Filetransfers	1968		
Data into dCache [MBytes]	728012.6	Data rate into dCache	202.2 MB/s
Data out of dCache [MBytes]	2601728.1	Data rate out of dCache	722.7 MB/s
Data written to tape[MBytes]	165437.2	Average tape write speed	46.0 MB/s
Data read from tape[MBytes]	38803.3	Average tape read speed	10.8 MB/s



Monitoring

VO	data on disk [TB]	files on disk	data on disk [TB] (disk-only pools)	files on disk (disk-only pools)	MB to tape	tape write speed [MB/s]	files to tape	MB from tape	tape read speed [MB/s]	files from tape	MB into dCache	rate in [MB/s]	MB out of dCache	rate out [MB/s]	total file transfers
atlas	339.7	3471141	241.3	3269145	0	0	0	0	0	0	22642.6	6.3	1613969.6	448.3	248
alice	46.4	58292	0	36	74894.5	20.8	92	0	0	0	0	0	0	0	0
cms	384.1	249106	5.7	11035	90542.7	25.2	69	38803.3	10.8	14	704704	195.8	916988	254.7	1367
lhcb	68.8	339513	46.7	291088	0	0	0	0	0	0	0	0	0	0	0
cdf	12.2	15393	0.2	365	0	0	0	0	0	0	0	0	45814.6	12.7	117
compass	22	48706	8.5	22609	0	0	0	0	0	0	0	0	24954.9	6.9	18
auger	3.7	20930	3.7	20930	0	0	0	0	0	0	665.1	0.2	0	0	9
dteam	0.1	10498	0.1	6608	0	0	0	0	0	0	0.9	0	0.8	0	175
ops	0	1661	0	1661	0	0	0	0	0	0	0.1	0	0.1	0	34



▶ DESY Tier-2 storage

- *dcache-se-atlas*: 161 TB
- *dcache-se-cms*: 170 TB
- additional ATLAS (150TB) and LHCb instances planned in Zeuthen

local access from Tier-2 and NAF computing resources

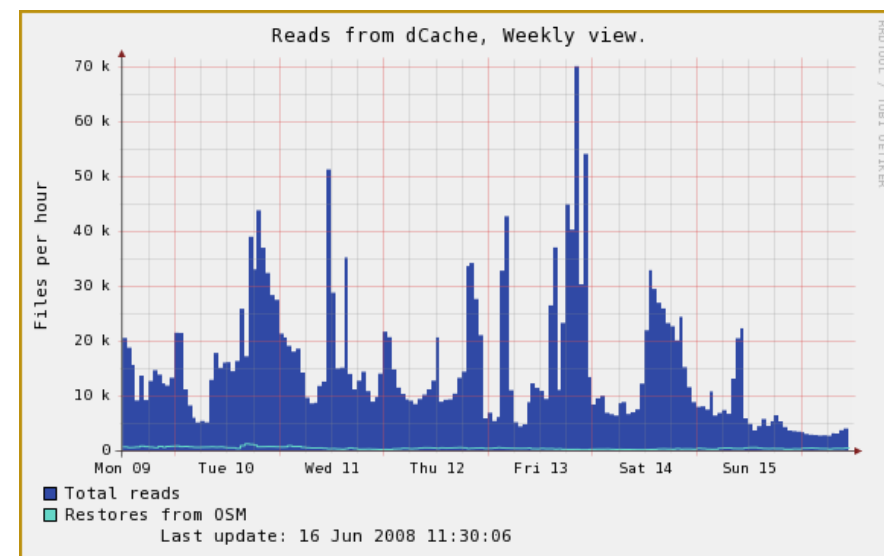
▶ VOs ILC, CALICE

- *srm-dcache* soon replaced by *dcache-se-desy*
- tape backend



▶ Hera dCache

- where it all started :)
- main instance used by H1, Zeus, Hermes, HeraB
- dedicated H1 resilient dCache





Current and future developments



Chimera namespace provider

The PNFS namespace provider is the expected performance bottleneck, when dCache scales up to the Petabyte-range with millions of file entries.

- ▶ solution: **Chimera** as the drop-in replacement of dCache's namespace
- ▶ some features:
 - dCache not bound to the limitations of NFS Ver. 2/3 anymore
 - clean, database-centric design
 - more requests/sec
 - more file entries
- ▶ PNFS or Chimera: your choice
 - migration procedure/tools available
 - some Tier-1s in the process of conversion





Access Control Lists (ACLs)

- ▶ fine-grained NFS4-based ACLs on
 - files, directories, (in the future also on SRM space tokens)

- ▶ migration phase
 - fallback to UNIX permissions where no ACLs are set

- ▶ setACL/getACL
 - possible via SSH-interface
 - OR
 - via NFS4 mount

- ▶ for now, ACLs available for the GsiFTP-door
 - other protocols will follow
 - will be part of dCache 1.8.0-16



The new Info system

- ▶ collects various information from your dCache
 - snapshot of the whole dCache instance
 - topology and health status of dCache cells, routing information
 - status of pools, SRM spaces
- ▶ information publishing to the grid
 - using full Glue Schema v1.3 (including SRM spaces)
 - we are active member in the Glue 2.0 specification group
- ▶ information publishing for local usage
 - all info XML encoded and served via HTTP
 - rich ssh-admin interface

again, this will come with version 1.8.0-16



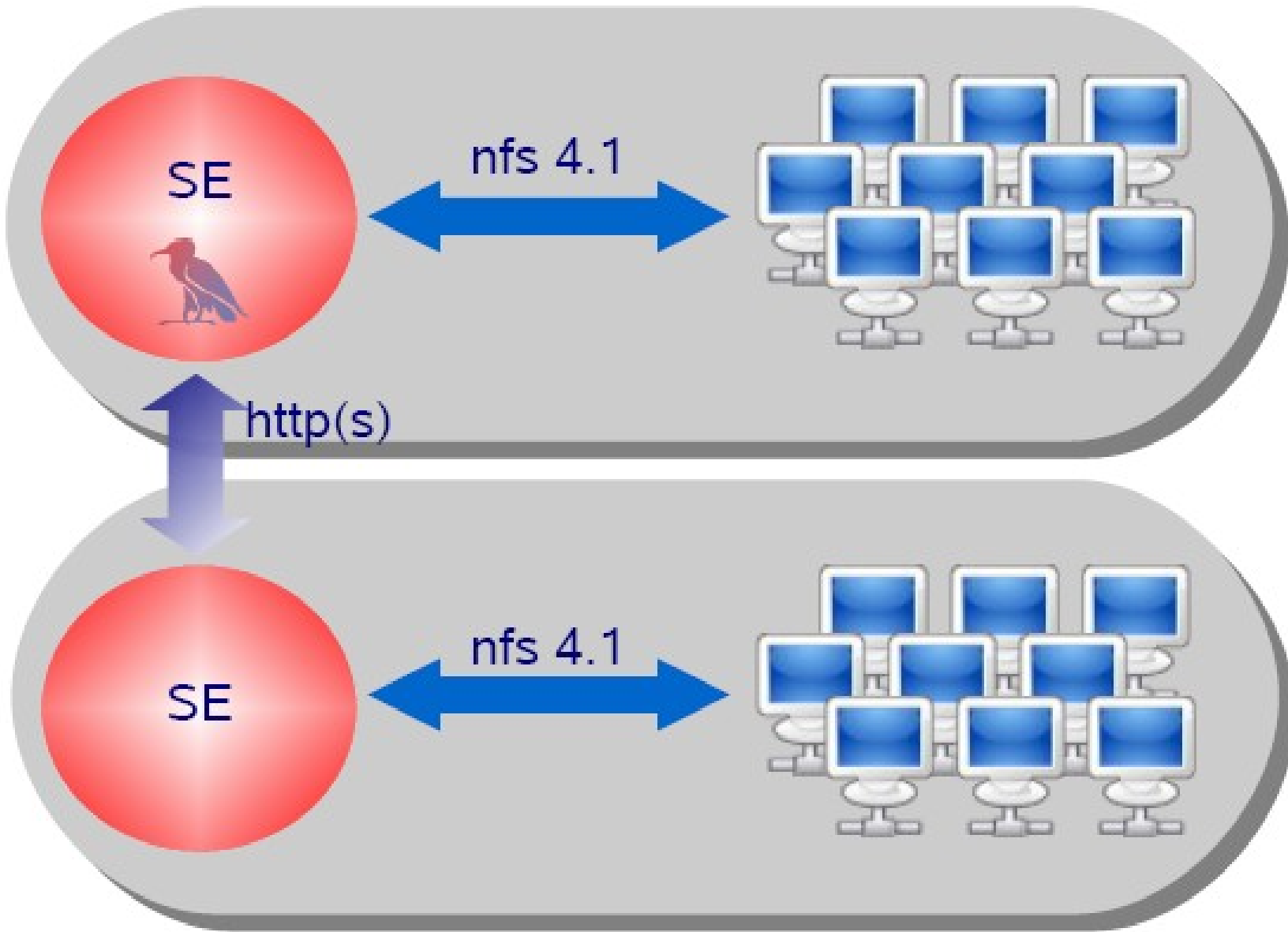
NFS 4.1 support

- ▶ we are actively involved in development of protocol spec and an implementation for dCache
- ▶ technical advantages
 - NFS 4.1 finally aware of **distributed data** (fit's dCache better)
 - clients will come for free (contained in all major OS)
 - GSS auth., ACL support, smart client caching
- ▶ timeline
 - prototype server in dCache 1.8.0-16 release
 - production grade by the end of '08
 - NFS 4.1 client in Linux kernel by the end of '08

*Leveraging industry standards will make
dCache more attractive for Non-HEP communities*



Goal: meeting industry standards





Thanks for your attention!

Contact:

www.dcache.org

Specific help for your installation:

suport@dcache.org

User Forum:

user-forum@dcache.org

Want to try out dCache without the hassle of configuring?

· Install VirtualBox on you Desktop and get the VM image

<http://trac.dcache.org/trac.cgi/wiki/dCacheToGo>