AstroGrid-D Meeting at MPE
14-15. November 2006
Garching

dCache

A scalable storage element and its usage in HEP

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Martin Radicke
Patrick Fuhrmann

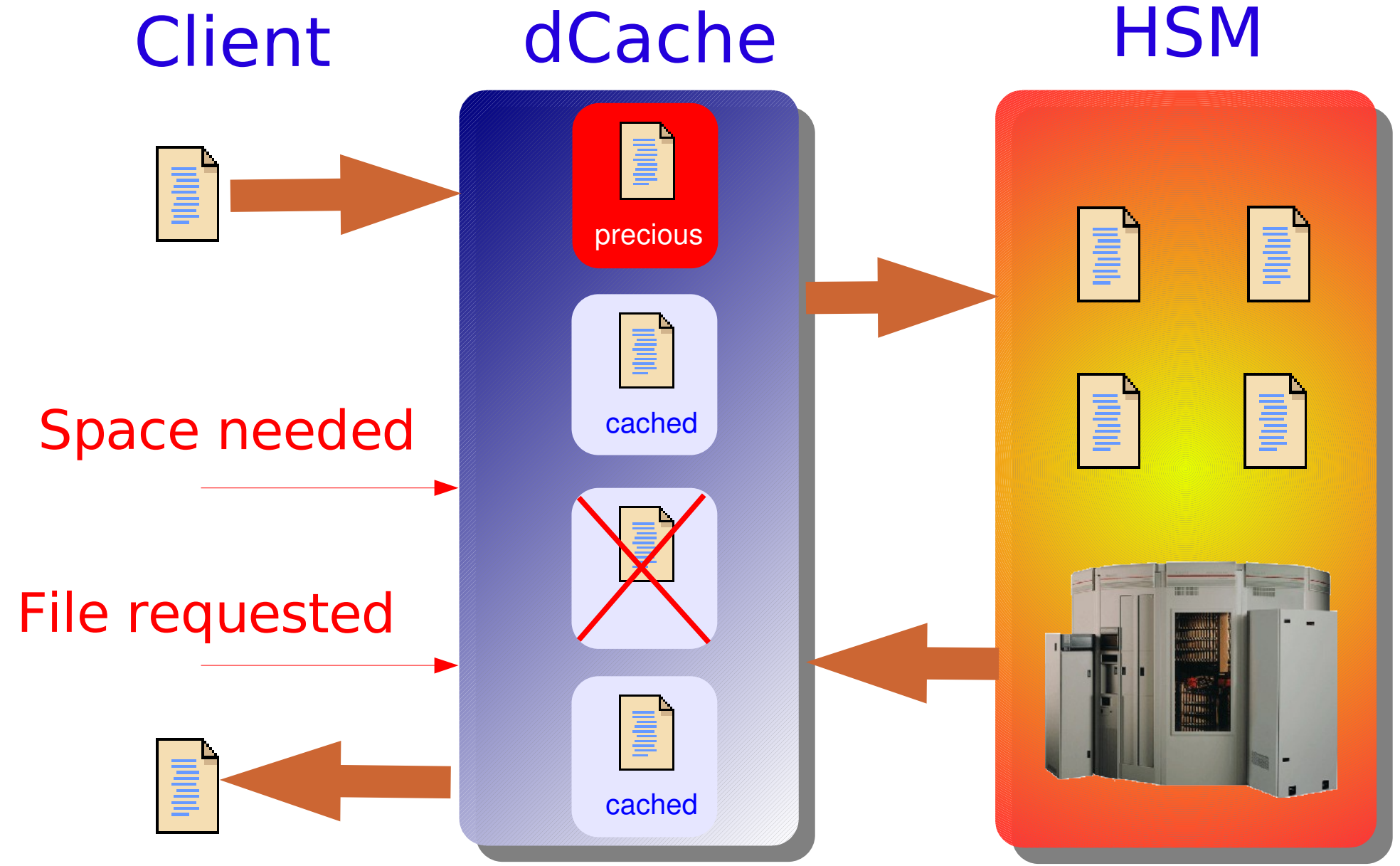
DESY

Introduction to dCache

Project overview

- joint venture between DESY and FERMI Lab
  - 10 FTE in total
- development/ in production since 2000/03
- MoUs with: LHC, OSG, D-Grid (HEPCG, DGI)
- DESY provides first line support
  - Trouble-Ticket-System, user forum mailing list, central documentation (1 FTE)
  - workshops
- Sourcecode available at www.dcache.org

The basic idea



What is dCache?

- dCache is managed storage
- storage element (SE)
  - full Storage Resource Manager (SRM) support
  - variety of data access protocols
    - local area: dCap, xRoot
    - wide area: GridFTP, (HTTP)
  - information providing: GIP (LCG), JClarens (OSG)
- frontend to Tertiary Storage Systems
  - supported: OSM, TSM, Enstore, HPSS

dCache key features

- combines hundreds of commodity disk servers to get a huge PetaByte-scale data store
- 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
- automatic file replication on high load (hotspot detection)

Insight dCache: components

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

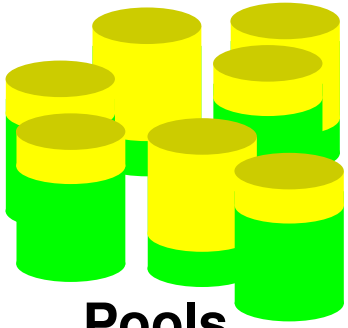
- provides single rooted namespace service
- one central instance
- can be viewed and modified via NFS 2/3 mount, FTP commands

SRM

- Storage Resource Manager
- identified by TURL
- selects Door based on load and agreed transfer protocol

Doors

- protocol-specific entry points to the file repository
- identified by SURL
- contacted by Clients to initiate file transfer



Pools

- diskserver holding (caching) 0..n copies of files known to dCache
- get selected and do the actual file transfers (called movers)
- migrating files to and from Tertiary Storage

Pool Selection & Load Balancing

- run by the Poolmanager (the heart of dCache)
- Select a **set of pools** which matches the following criteria
  - Protocol
  - Dataflow Direction (put, get, restore from HSM, pool2pool)
  - Directory Subtree
  - Client IP Address
- Out of these, select the “best” **target pool** with lowest cost
  - get request: cost = cpu load (number of mover)
  - put request: cost = free space + cpu load

LAN filetransfer protocols

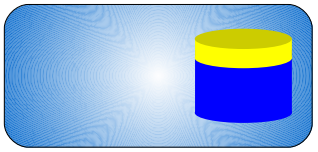
- Transparent access from within ROOT toolkit
  - dCache behaves like a xrootd server cluster
  - protocol implementation makes full use of dCache core features (load balancing, HSM backend)
  - token-based authorization
- DCap – the native dCache protocol
  - advanced tuning options
  - New: passive mode (to get access from behind firewalls/NAT)
  - authenticated flavour available (gsiDCap)

Storage Resource Manager (SRM)

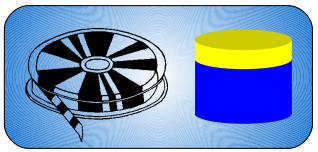
- an interface for standardized access of SEs
- main functions:
  - prepares transfers (resolves SURL->TURL)
  - negotiates protocols (in theory)
  - initiates pre-staging
  - provides directory functions (e.g. ls)
- clients
  - Command Line Tool, Generic File Access Library, File Transfer Service

SRM v2.2

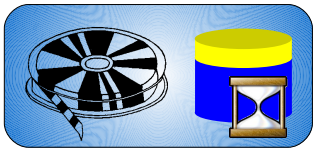
- space reservation (space token)
  - can be set per SE and Virtual Organisation
- storage classes



no HSM, or file is never written to HSM



file is written to HSM but kept on disk forever



file is written to HSM and kept on disk for a certain time (Pinning)

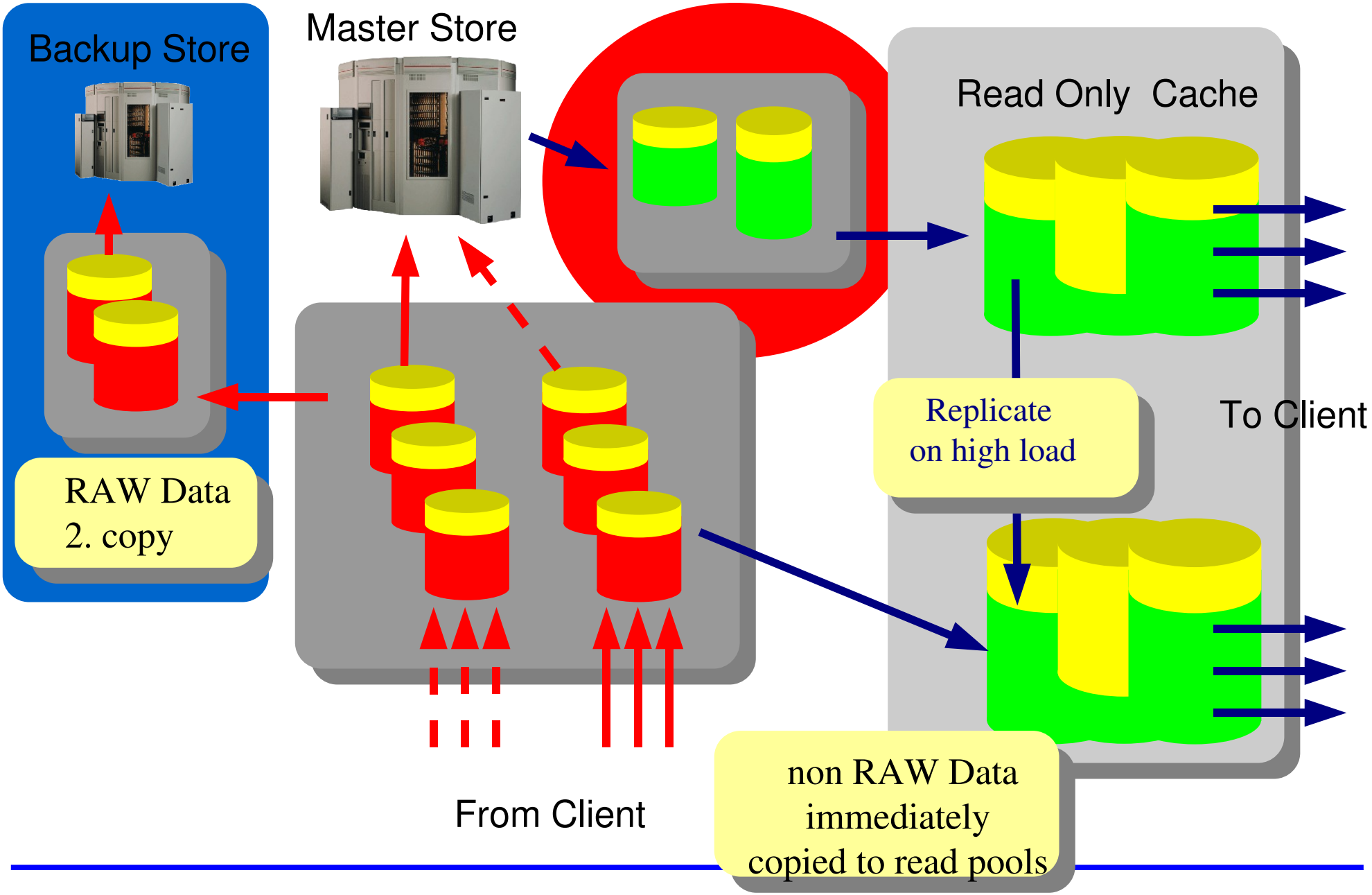
- full SRM v2.2 implementation for dCache is under development

dCache Scaling

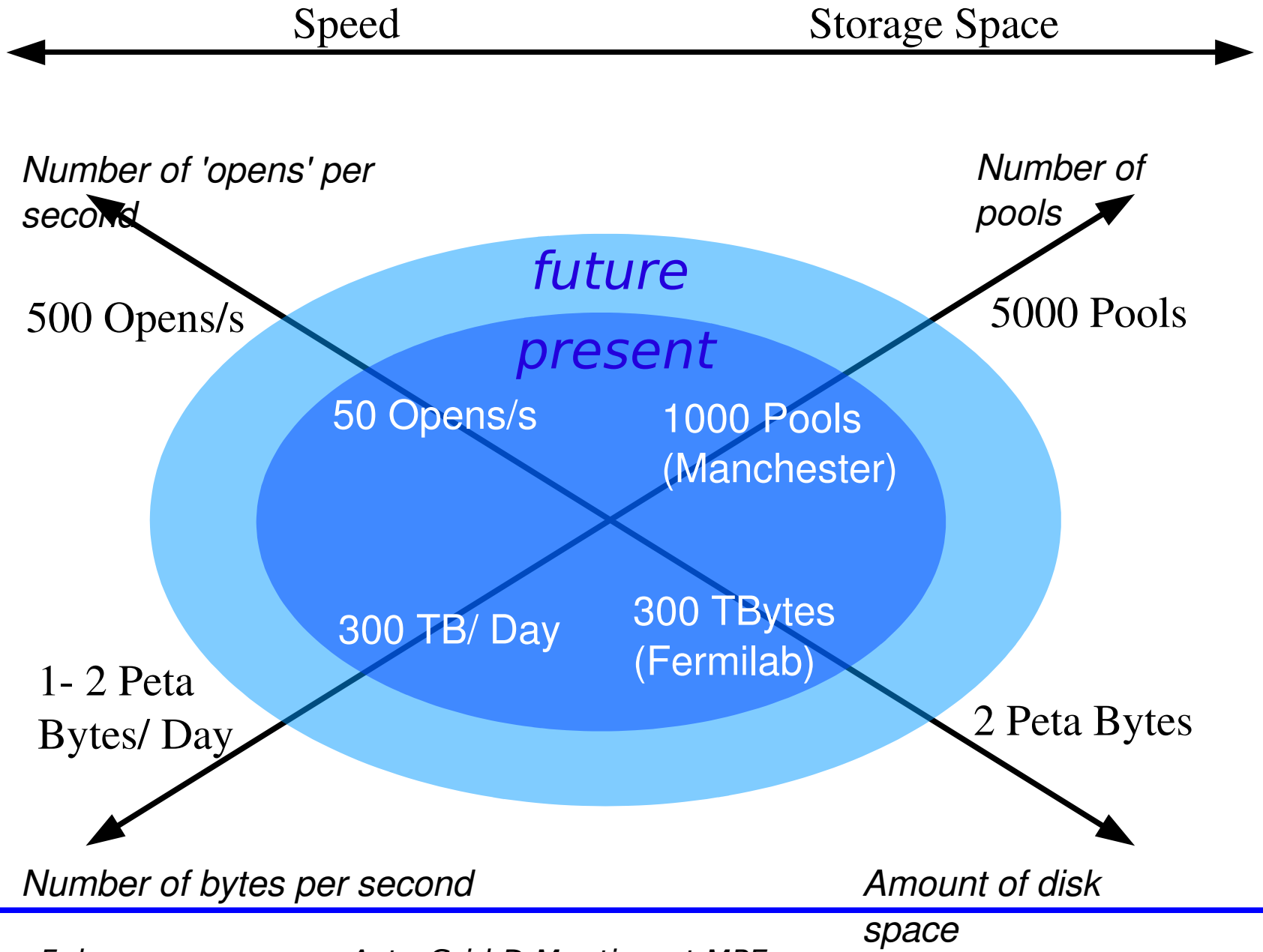
Optimizing internal dataflow

- Smart handling of bunch requests by a fast pool selection unit
- Intelligent HSM backend migration
  - Centrally controlled mechanism to optimize HSM interaction
  - alternate flushing strategy to reduce tape mounting times
- File Hopping
  - Automatic data set replication on hot spot detection
  - allows write-only pools: replication on demand or on arrival

Smart flushing & file hopping

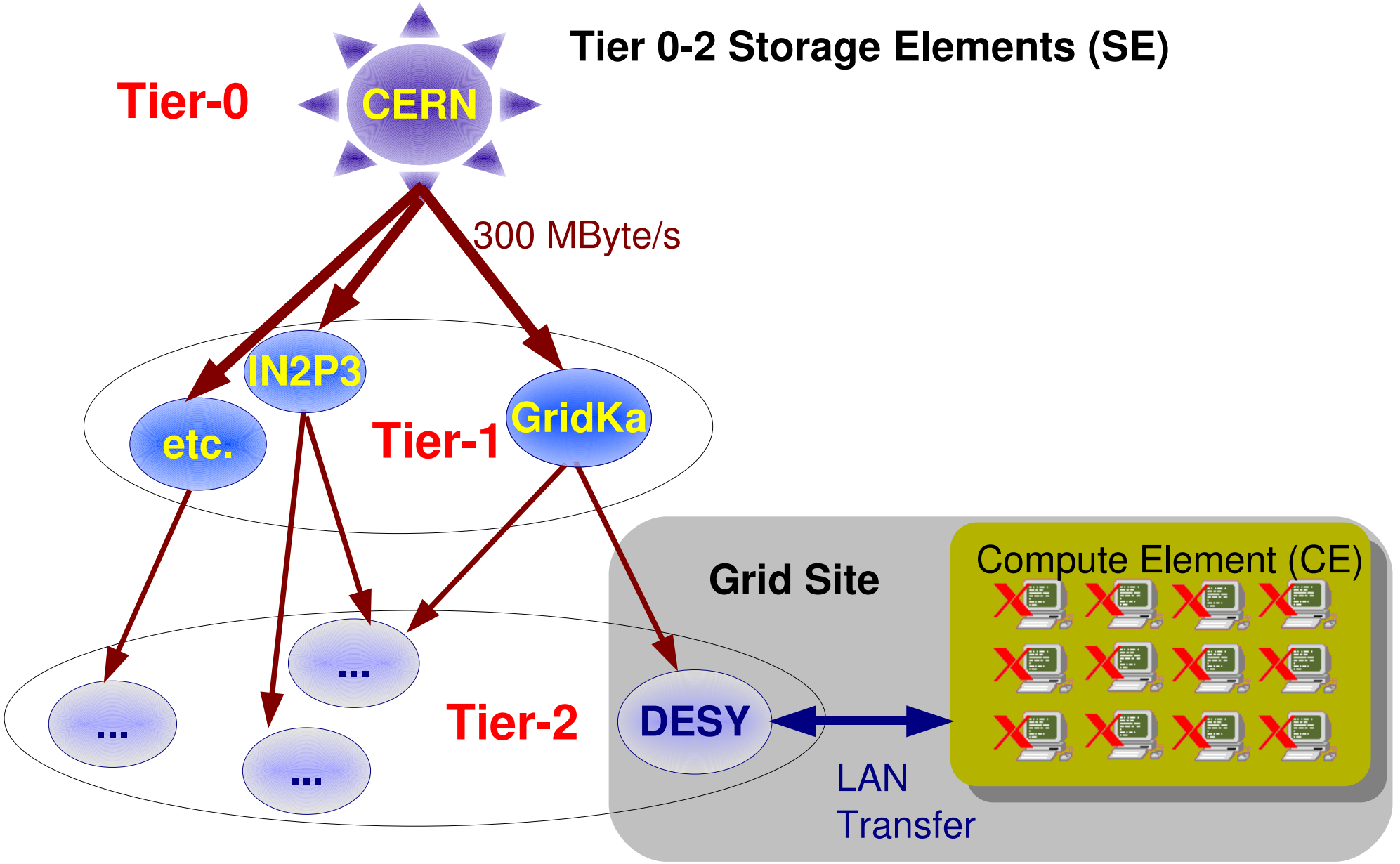


SE scaling in the near future

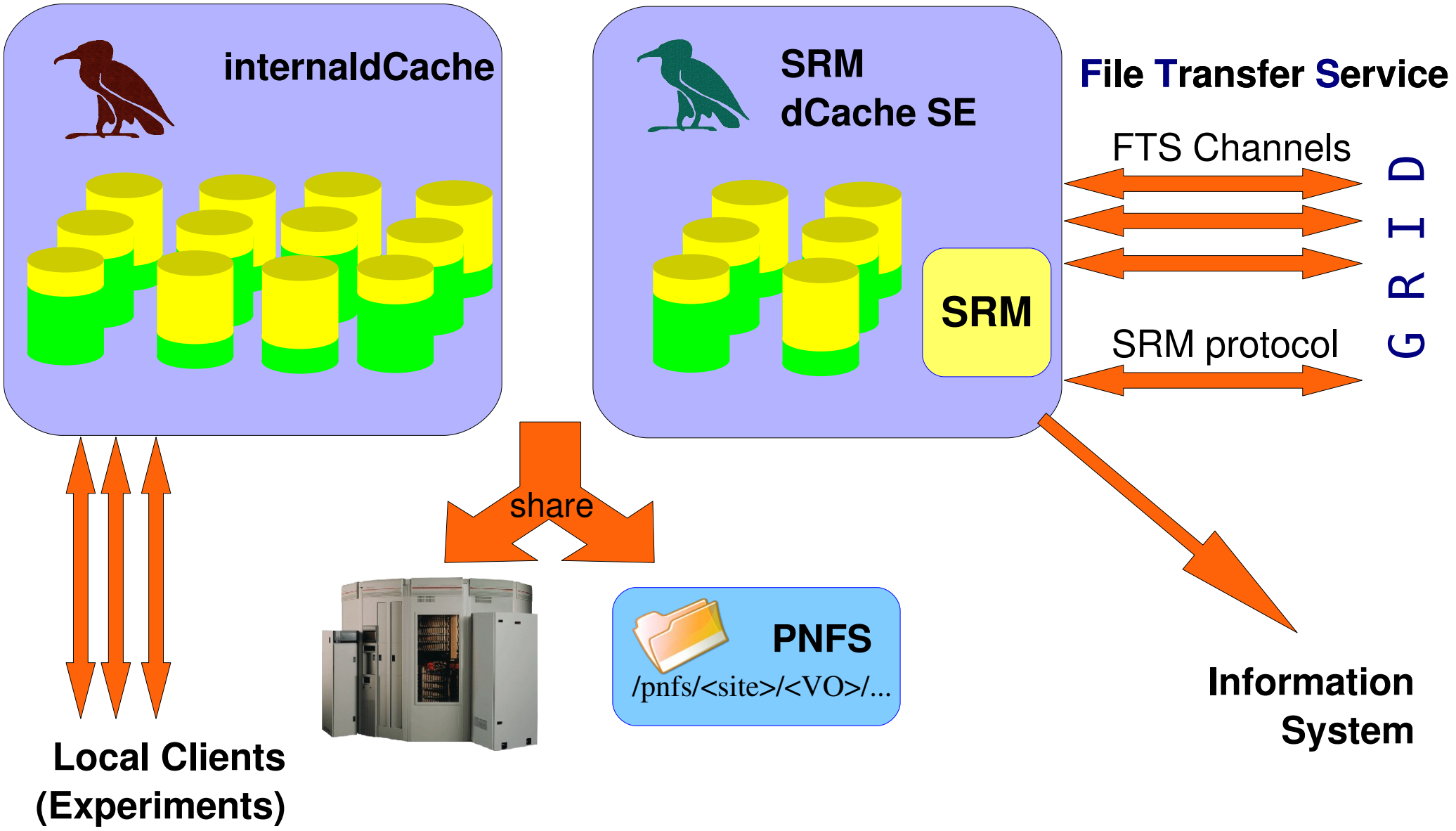


The dCache SE in HEP

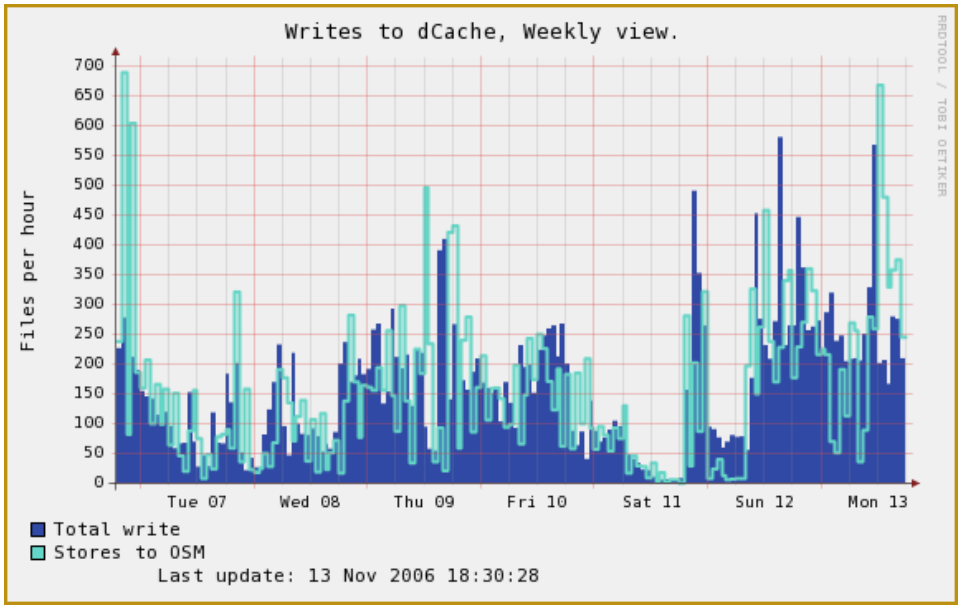
LHC Data Grid Schema`



dCache@DESY

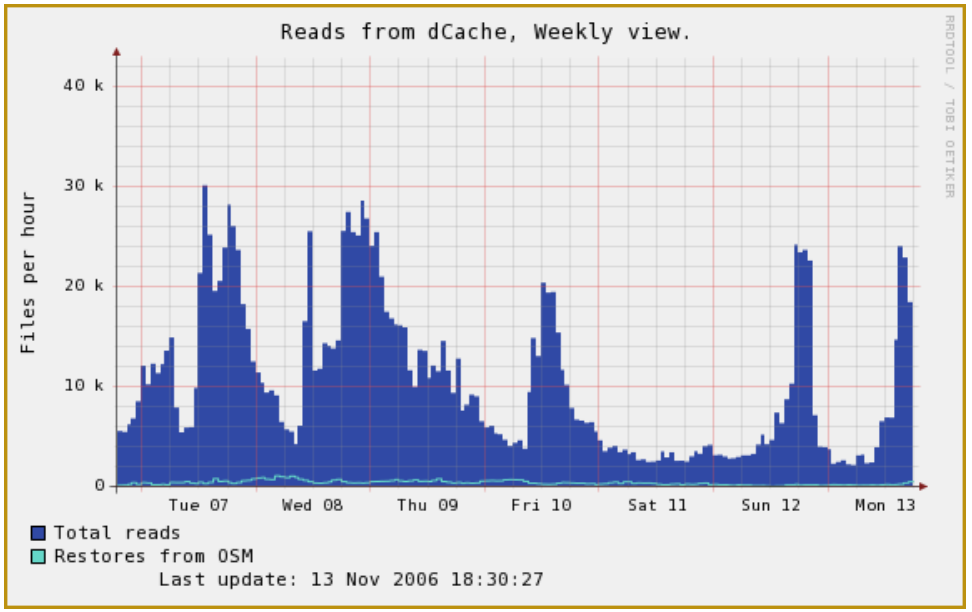

internaldCache
SRM
dCache SE
SRM
File Transfer Service
FTS Channels
SRM protocol
G R I D
share
PNFS
/pnfs/<site>/<VO>/...
Information
System
Local Clients
(Experiments)

Throughput at DESY (central dCache)



Almost all datasets written into dCache are migrated to magnetic tape (raw data)

Most datasets requested for read are still on disk
→ high cache efficiency



Installation
&
Configuration

Deployment

- latest production version: dCache 1.7.0
  - Binaries for Linux, Solaris, (WinXP)
  - available from http://www.dcache.org/
    - direct download of RPMs
    - APT/YUM repositories
    - source code
- manual installation (the standard method)
- new: automatic installation/upgrade via YAIM
  - a full single-host dCache instance in 10min!!
  - can handle more complex setups (multi-host)

Administration and Monitoring

- feature-rich ssh admin interface
- GUI-application
  - full commandset available
  - Topology browser
  - visualises cost tuning and central flush management
- webinterface for monitoring
  - status of each service
  - queue lengths
  - Pool-and HSM- usage
  - dataflow rules
- billing database for customized statistics

- collaboration with DGI
  - providing dCache-SEs as Core-D-Grid resources
  - accessible via VO 'dgtest'
- existing installations
  - FZ Jülich (Terabyte-scale, Tape backend)
  - RWTH Aachen
  - FZK Karlsruhe
  - DESY Hamburg

Future development

- full SRM 2.2-compliance
  - expected in Spring 2007
- Extended Information System (HEPCG)
  - How long does it take to get a file ready for i/o ?
- Chimera (Improved file system engine)
  - better performance for large-scale installations
  - Acl's, Quotas
- nfs4.1 (including data transport)
- Improved HSM connectivity (central staging)

2nd dCache Workshop

18-19 January 2007
DESY, Hamburg

Agenda:
* The SRM 2.2 implementation
* Space Tokens and Storage spaces
* dCache operational issues
* Yaim (Quattor) Installation
* Meet the FERMI and DESY developers
* Reports from some of the Tier I sites

Register at: https://indico.desy.de/conferenceDisplay.py?confId=138

Contact:	www.dcache.org
Specific help for your installation or your customized dCache instance:	suport@dcache.org
User Forum:	user-forum@dcache.org