HEP-CG Workshop
30.11 - 1.12. 2006
Wuppertal

Update

# dCache

## A scalable storage element

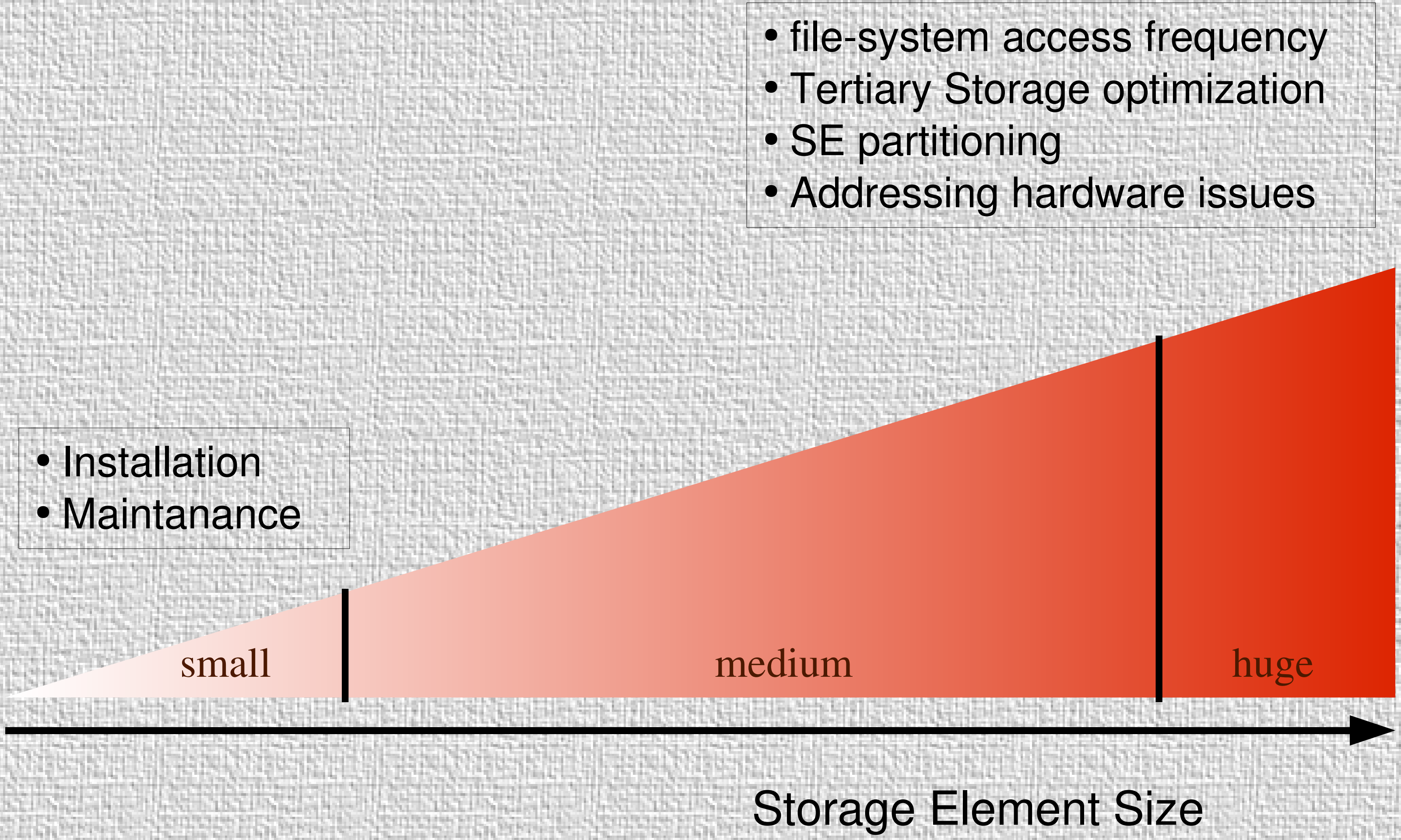
Martin Radicke

Patrick Fuhrmann

DESY

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

# Scaling issues



- file-system access frequency
- Tertiary Storage optimization
- SE partitioning
- Addressing hardware issues

- Installation
- Maintanance

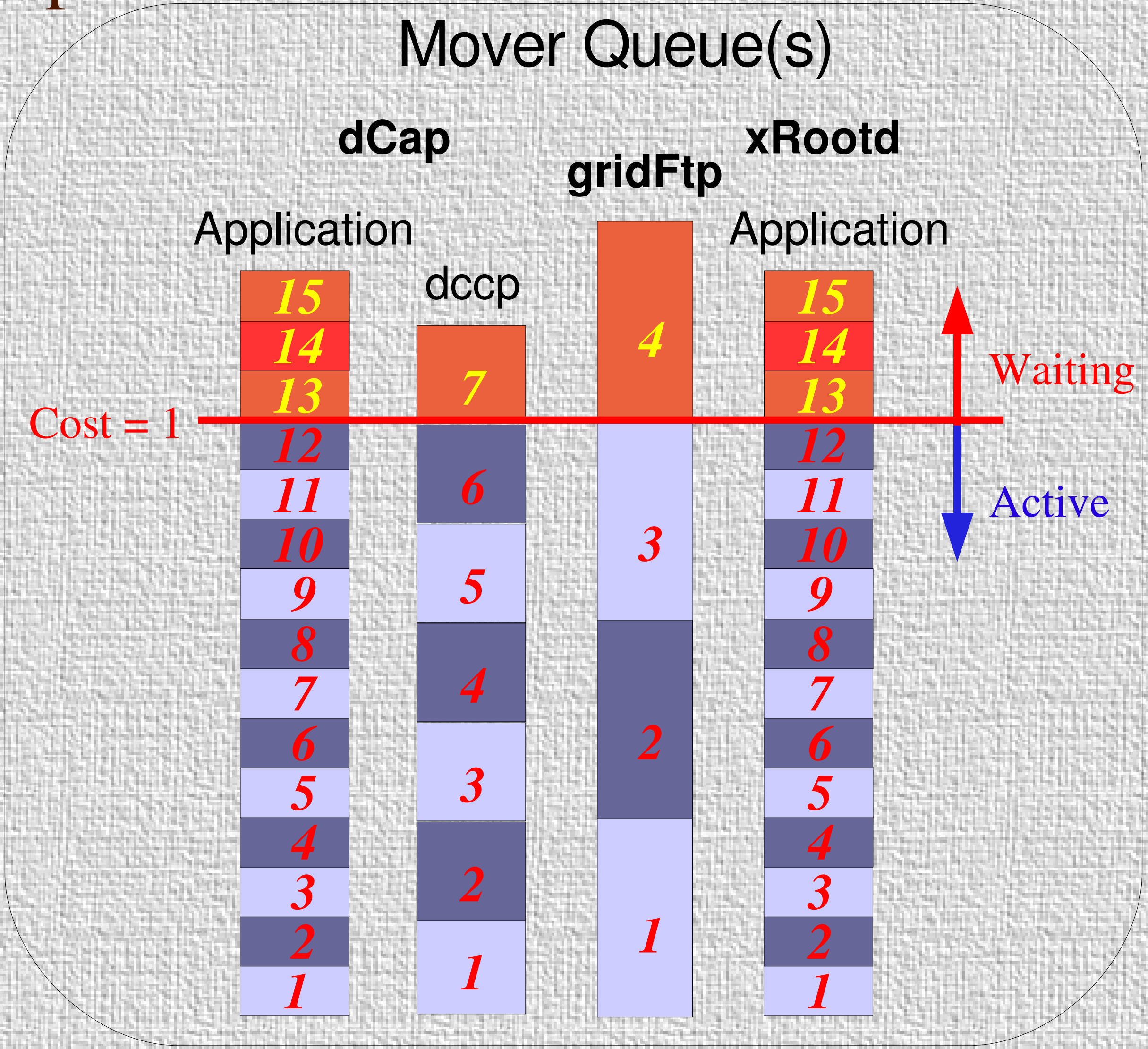
Storage Element Size

# Small scale

- YAIM-based installation
  - full SRM/dCache-SE in 20min for a single-host instance
    - including SRM, InfoProvider, multiple VOs
  - provides good starting point for future growing
  - can be rerun to apply changes
    - distributing components on multiple hosts, adding disk space or VOs
- most of the advanced dCache tuning options now available in one central setup file
  - e.g. port ranges, authentitication policy, protocol parameters, ...

# Mid scale (T 2.1)

- multiple I/O-queues on pools
  - handle slow and fast transfers differently to reflect usage patterns



# Mid scale (cont'd)

- File hopping (data set replication)
  - incoming data sets often go to special write pools (write-only cache)
    - replication to read-pools on arrival OR on read request
  - automatic replication on hot spot detection
- Centrally managed flushing
  - only a controlled amount of streams go to the HSM backend
  - alternate flushing to optimize HSM access (improves disk performance and therefore tape throughput)

# Huge scale (T 2.1)

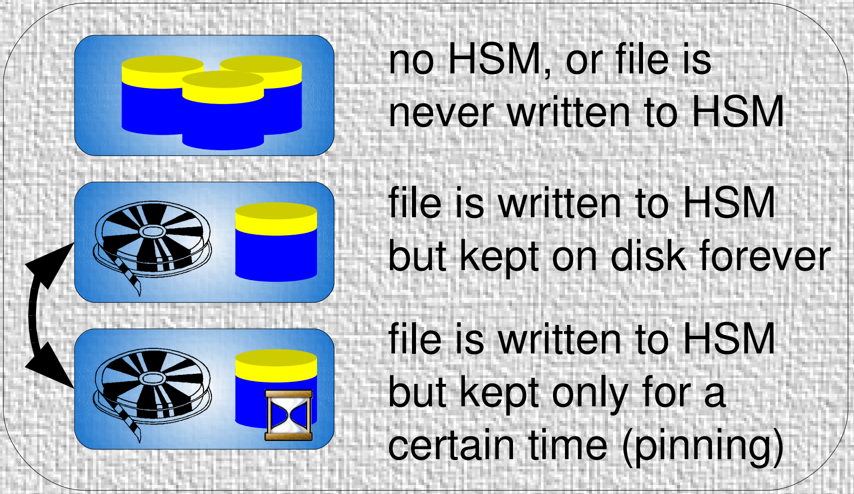
- new namespace- and metadata engine: CHIMERA
    - replacement of PNFS service
    - better performance
        - only a few big tables ➡ RDBMs optimise on this
        - Metadata not stored in BLOBs anymore ➡ direct queries
        - direct dCache ↔ NameSpace communication (avoids NFS bottleneck)
        - NFS v2/3 "view" still provided for legacy clients
    - all RDMSs with JDBC drivers supported
        - sucessfully tested with Oracle, MySQL, PostgreSQL
        - scaling achieved through DB-clustering
    - can be extended to manage arbitrary metadata
    - ACL support in preparation
    - in beta-testing phase, expected in Spring 2007

# product improvements(T2.1)

- extended certificates (voms-proxy-init)
  - supporting  VOs and Roles
    - one DN in several VOs, several roles for one DN, DN-based exclude list
- dCache partitioning to reflect usage scenarios
  - pool selection mechanism can be tuned for each partition differently
    - e.g. write request distribution: empty space filled first vs. smooth overall load
- more on transfer protocols
  - dCap – the native dCache protocol
    - passive mode was added for access from behind firewalls/NAT
  - xrootd officially supported
    - tokenbased authorization (ALICE security model)
    - gsi authentication coming soon

# SRM 2.2 (T 2.1)

- makes storage manageable from the client side
- Space Reservation (Space Token)
  - can be set per SE and VO
- Storage Classes
  - define data quality in terms of reliability and access latency
  - three classes and transitions between them agreed



no HSM, or file is never written to HSM

file is written to HSM but kept on disk forever

file is written to HSM but kept only for a certain time (pinning)

- dCache implementation status
  - under heavy evaluation, expected to be available in spring 2007

# GLUE Schema

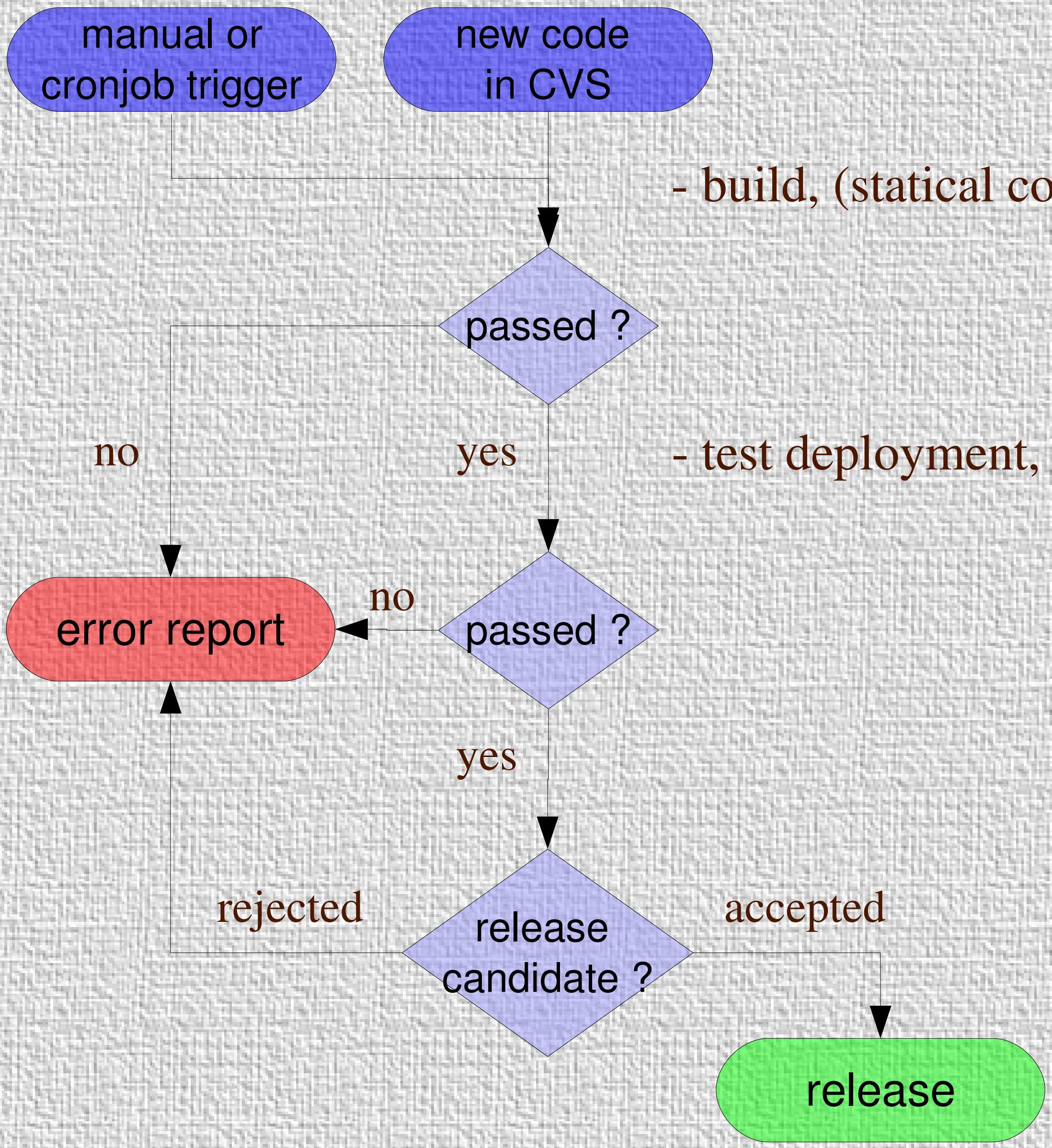
- current version supported by Information Provider: v1.2
  - e.g. available/used space per Storage Element (SE) and per VO
- implementation of new version (v1.3)

  begins as soon as it is agreed
  - online/nearline space of the entire SE
    - used/available space deprecated
  - new StorageArea (SA) definition to support SRM v2.2 protocol
    - describes a portion of physical space, spans different kinds of storage devices
    - each instance implements one of the SRM v2.2 storage classes
    - only one VO per SA allowed in order to allow dynamic space reservation (in LCG)
    - one VO can own multiple SAs on each Storage Element

# Agreement with LCG

- Availability of dCache software
  - stable release in CERN apt-repository as part of gLite
  - most recent release in DESY apt-repository
    - allows much faster update cycles
- LCG deployment group
  - performs dCache installation verification
  - performs gLite interoperability verification against recent dCache instance at DESY
  - dCache contact person at CERN

# Automatic build & test



- build, (statical code analysis), packaging (T 1.5)

- test deployment, automatic run of testsuite (T 1.4+2.5)

`

- final decision made manually

- published to apt/yum repository (T 2.6)
  and to website

- changelog, announcement

# Support (T 2.6)

- 1 FTE responsible
    - trouble ticket system
    - user forum
    - documentation (soon)
    - contact to LCG and Grid PP
    - workshops
- current release: dCache 1.7.0
    - migration paths from older versions available
- source code available at www.dcache.org
    - special license
    - support restricted to official binaries

# Administration Tools (T 2.4)

- Java/Python interface
  - health monitoring, programmatic access to runtime parameters
  - basis to build more sophisticated tools upon
- Improved GUI tool
  - cost overview & tuning
  - execute actions on entire poolgroups
  - central flush management
- statistics module
- 3rd-party plugins for Nagios and Monami

# Co-Scheduling

- Extended Information Provider (EIS)
  - interface to retrieve file-specific status in order to make dCache SE a more planable resource
  - file online/nearline?
  - time to get file ready for transfer (restage time)
    - new component: prediction engine
    - current implementation makes use of HSM access history statistics and derives simple predictions
    - more sophisticated algorithm needed
    - requires real HSM interactions for testing

# Contact

dCache, the Book

*www.dCache.org*

need specific help for your installation or help in designing your dCache instance.

*support@dCache.org*

dCache user forum

*user-forum@dCache.org*