dCache workshop at DESY

- Scope and participants
- Program
- Summaries of talks and discussion
- Conclusion

Scope and participants

- Scope
  - Analysis of dCache configuration and performance, in particular w.r.t. SC3 experience → how to get to an optimal setup satisfying LCG requirements
- Participants
  - Mostly sysadmins of SC3 and other dCache sites
    - DESY, FNAL, gridKa/FZK, IN2P3 Lyon, RAL, SARA, UK T2, BNL
  - dCache developers
  - CERN SC3 and deployment delegation
- Dates
  - Aug. 30 – Sep. 1, 2005
- URL
  - http://www.dcache.org
  - Original presentations under “documentation”

Program

- Patrick Fuhrmann (DESY) - “Insight dCache”
  - showing expert "tricks and tips" in response to questions and concerns collected prior to the workshop
- Timur Perelmutov (FNAL) - “dCache SRM”
- James Casey (CERN) - “dCache in SC3”
- SC3 site reports
  - gridKa (FZK), IN2P3 Lyon, RAL, SARA, UK T2, BNL
- Discussion
- Maarten Litmaath (CERN) - “dCache SE in LCG-2”
- Patrick Fuhrmann - demo of dCache admin GUI
- Martin Gasthuber (DESY) - “Storage Task Force”

Insight dCache summary I

- dCache collaboration and code organization
- Comprehensive admin guide ("The Book") by Matthias de Riese (DESY)
  - http://www.dcache.org/manuals/Book/
- Illustration of architecture
  - Domains, cells, and their startup scripts
  - Example interactions with domains and cells
- Site setups
  - Head (admin) node, pool nodes, door nodes
  - Admin node can be split
    - E.g. SRM can be put on separate node
  - Doors can be put on pool nodes with inbound access (see later)
  - PNFS default and customized setups

Insight dCache summary II

- Pool Manager
  - Pool Selection Unit finds allowed pools, Cost Manager find best, based on client IP, storage class, I/O direction
- Reads can use cheap disks, writes should use best disks in the system
- Pool cost has 2 components
  - Mover cost increases with active and waiting movers
    - Mover queues for fast gridftp transfers vs. slow DCAP transfers
  - Space cost increases as LRU file is younger
  - Total cost linear combination of both (can be tuned)
    - Space cost factor is zero for reads
- Thresholds for pool-to-pool transfers (load-balancing)
  - Move files to cheaper pools
- dCache version 1.6.5 had a bug in the Cost Manager configuration, degrading SC3 performance → work-around provided
- Admin GUI very helpful to see what is going on with pools etc.

Insight dCache summary III

- Resilient Manager
  - Control nr. of replicas
  - Control pool draining
- Internal Copy Manager
  - Copy data sets to given pool
- PNFS development
  - NFSv2 → NFSv3
  - GDBM → Postgres
    - No 2 GB max. size
    - Better performance
    - Continuous backup
- VOMS authentication prototype
  - No groups/roles yet

dCache SRM summary I

- SRM collaboration
- SRM motivation
  - Reservation and scheduling of heterogeneous storage
- SRM role in data life cycle
- v1.1 (current) vs. v2.1 features
- SRM-dCache communication
- Network flows → firewall configuration
  - GET
  - PUT
  - SRMCP pull mode
  - SRMCP push mode

dCache SRM summary II

- Status of FNAL SRM implementation
  - Data Transfer Functions (get, put and copy)
  - Load balancing, throttling, fairness
  - Scalable replication mechanism via gridftp
  - Automatic directory creation
  - Fault tolerance via transfer request DB and retries
  - Standalone SRM interface
  - SRM-Storage interface to UNIX file system
  - Implicit space management

dCache SRM summary III

- FNAL SRM plans
  - Full implementation of SRM Version 2.1 interface
  - Explicit Space Management
  - Support for at least Volatile and Permanent space types
  - Directory and Permission functions
  - Use of Lambda Station Interface for on-demand optical path allocation
  - Monitoring, Administration and Accounting interfaces

dCache in SC3 summary

- Most SC3 sites use dCache
- Performance varied a lot between sites
  - Transatlantic vs. “short-haul” networks
  - Number of streams
    - Sometimes more is better, sometimes not
  - Timeouts → retries
    - Movers often not cleaned up
  - Low network utilization
  - Kernel tuning
    - TCP and disk I/O buffer sizes
    - Ext3 vs. GPFS
- Related issues with Castor and DPM
- GridFTP performance markers would help
  - dCache developers agreed to implement them

Site report summaries I

- gridKa (FZK)
  - Inbound vs. outbound rates
  - Some multi-homed nodes had wrong NICs used for transfers
    - Explicit configuration needed
  - File delete overhead in ext3 (bad) vs. GPFS (good)
- IN2P3 Lyon
  - HPSS back-end via RFIO
    - Pool directory currently has to be world-writable
    - Issues with load-balancing, timeouts
      - But all files were stored in single HSM directory per VO
  - Would like access to sources
    - Private arrangement has been made
    - Sources not publicly accessible before CHEP '06

Site report summaries II

- RAL
  - Postgres DB on separate node for CPU load
  - Second SRM node for tape access
    - Different file lifetimes
    - LCG information system only allows one storage root per VO per SE
  - Pool disk NFS-mounted → NFS hanging
  - Use PNFS tags to send files to VO's pool group (as documented)
  - Problems due to lcg-gt existence checks without lcg-sd cleanup
    - Blocks transfer slot (default 24h)
  - FTS gave up on transfers without informing SRM
  - Pinmanager hangs → fixed
  - Java canonical hostname caching for castorgrid.cern.ch → fixed
  - Postgres slowing down due to lack of cleanup → fixed

Site report summaries III

- SARA
  - DMF file system front-end for HSM back-end
  - Using older dCache version (1.2.2-7-3)
  - Default nr. of I/O movers 100 → very high load
    - Set to ~5
  - Default heartbeat 120 sec → mistakenly suspected to lead to poorly balanced pools
    - Set to 10 as a test, but this parameter should not be decreased below 120
  - Nr. of gridftp streams
    - Globus gridftp server: 1-2 for optimal performance, 50 MB/s
    - dCache gridftp server: 1.6 MB/s per stream, used 10 streams (default)
  - Kernel VM tuning
  - SRM PUT → gridftp door on pool node → internal transfer to other node
    - Gridftp door selected independently of pool node → extra CPU+network load
      - Internal transfer can use second NIC
      - Hardware budget!
    - Fix requires major architectural change
      - GridFTP v2 X-mode can help
  - Timeouts → Postgres DB cleanup → fixed in later versions

Site report summaries IV

- UK T2
  - Edinburgh, Lancaster, Imperial College in SC3
    - Manchester, Glasgow, RAL-PP also installed dCache
  - Support group
    - http://wiki.gridpp.ac.uk/wiki/Main_Page
  - Easy to install with YAIM
  - Difficult to manage, maintain, configure
    - Improving also thanks to the workshop!
  - Log files difficult to understand
  - Script was needed to drain pools → fixed in upcoming version
  - SRM startup race condition
  - Firewall configuration
  - LCG information system needs tweaking
  - Can migrate Classic SE to dCache?
    - Not really
  - Unsupported OS on disk servers
  - Default 10kB writes → should be configurable
    - Comes from Globus
  - Excessive Java sockets in CLOSE_WAIT on pool nodes
  - Random SRM failures not understood

Site report summaries V

- BNL
  - dCache in production system since Nov. 2004
  - HPSS back-end
    - Oak Ridge Batch System optimizes prestaging
  - 322 farm WNs also act as read pool nodes!
  - 8 dedicated write pool nodes
    - XFS instead of ext3
  - 4 dedicated door nodes, PNFS + core services node, SRM node
  - 82 TB in data sets up to now
  - System successfully used for Atlas “Rome” production
  - USAtlas T2 sites to deploy dCache
  - Centralized PNFS potential bottleneck → being worked on
  - Network I/O bottleneck?
  - Pinmanager crashes → fixed
  - FTS does not support srmcopy yet → gridftp door bottleneck
  - Client hangs when pool node crashes during transfer

Discussion

- Single points of failure
  - PNFS, SRM, PoolManager, ...
- Gridftp server could demand transfer was requested via SRM
  - Avoid bypassing of load-balancing, fair-share, etc.
- How to schedule nr. of TURLs per VO/user?
- How to map a file to a grid user DN?
  - Accounting, auditing
  - To be added to billing file
- How to tie PNFS directories to database instances?
  - Static mapping, but less of an issue with Postgres
- Srmcp → gridftp should default to PASV mode
- Documentation
  - How-to for namespace layout, tie VO ←→ directories, tags

Sites in LCG-2 today

```
$ ldapsearch -x -h lcg-bdii.cern.ch:2170 -b o=grid | grep -c
    '^GlueSARoot.*/pnfs/'
97
$ ldapsearch -x -h lcg-bdii.cern.ch:2170 -b o=grid | grep
    '^GlueSARoot.*/pnfs/' |  sed 's-.*/pnfs/--;s-/.*--' | sort -u
cern.ch
desy.de
ft.uam.es
gridpp.rl.ac.uk
gsi.de
ifh.de
itep.ru
pp.rl.ac.uk
tier2.hep.man.ac.uk
zam.kfa-juelich.de
zib.de
```

...plus Fermilab and other sites that do not expose /pnfs...

Compatibility with other MW

- Note to developers and site admins:
  - A dCache SE must be usable as a standard LCG-2 SE
- That means:
  - Standard tools like lcg-cr should work
- Which means:
  - SE must correctly appear in information system
  - SE server code must be able to handle client code used in LCG-2, *even if the current client code is badly behaved* (we cannot fix the client code and upgrade the whole grid overnight)
    - Example: lcg-cr currently does not set file size in SRM put request
      - Bad for space reservation, but default can be used
      - Will be fixed for lcg-cr and friends
      - GFAL cannot set the file size, because POSIX open() cannot

dCache SE configuration for LCG-2

- Info provider
  - Currently hack provided by IT/GD
  - Native version being worked on at INFN Bari
- YAIM configuration
  - Maintained by GridPP (Jiri Mencak)
  - Should be extended to support pool selection parameters etc.
  - Should not configure dCache by default
    - Preserve existing configuration → needed to incorporate SC3 SEs into LCG-2
    - Admin can explicitly enable YAIM function in site-info.def

Storage Task Force summary

- Members
- Context
  - HEPiX, GDB
- Scope
  - What hardware for which profile, per Tier, per time?
    - Computing models
    - Data volumes
    - Access patterns
    - Security
  - Consider current technologies, prices per region
    - Trend analysis
    - Disk, tape, networks
  - Formulate plan for timely implementation of storage required
- Report at Oct. HEPiX at SLAC

Conclusion

- dCache workshop at DESY quite successful
  - Community of developers, site admins, etc. is forming
    - Social events were also appreciated!
  - Knowledge to be shared through user forum
    - mailing list, Wiki
  - Many issues discussed
    - Immediate solutions provided for “easy” problems
    - Feedback provided for near- and long-term development
  - Sites overall positive about dCache future
- Future workshops expected 1-2 times/year
  - Possibly in conjunction with LCG Operations workshop, CHEP, HEPiX