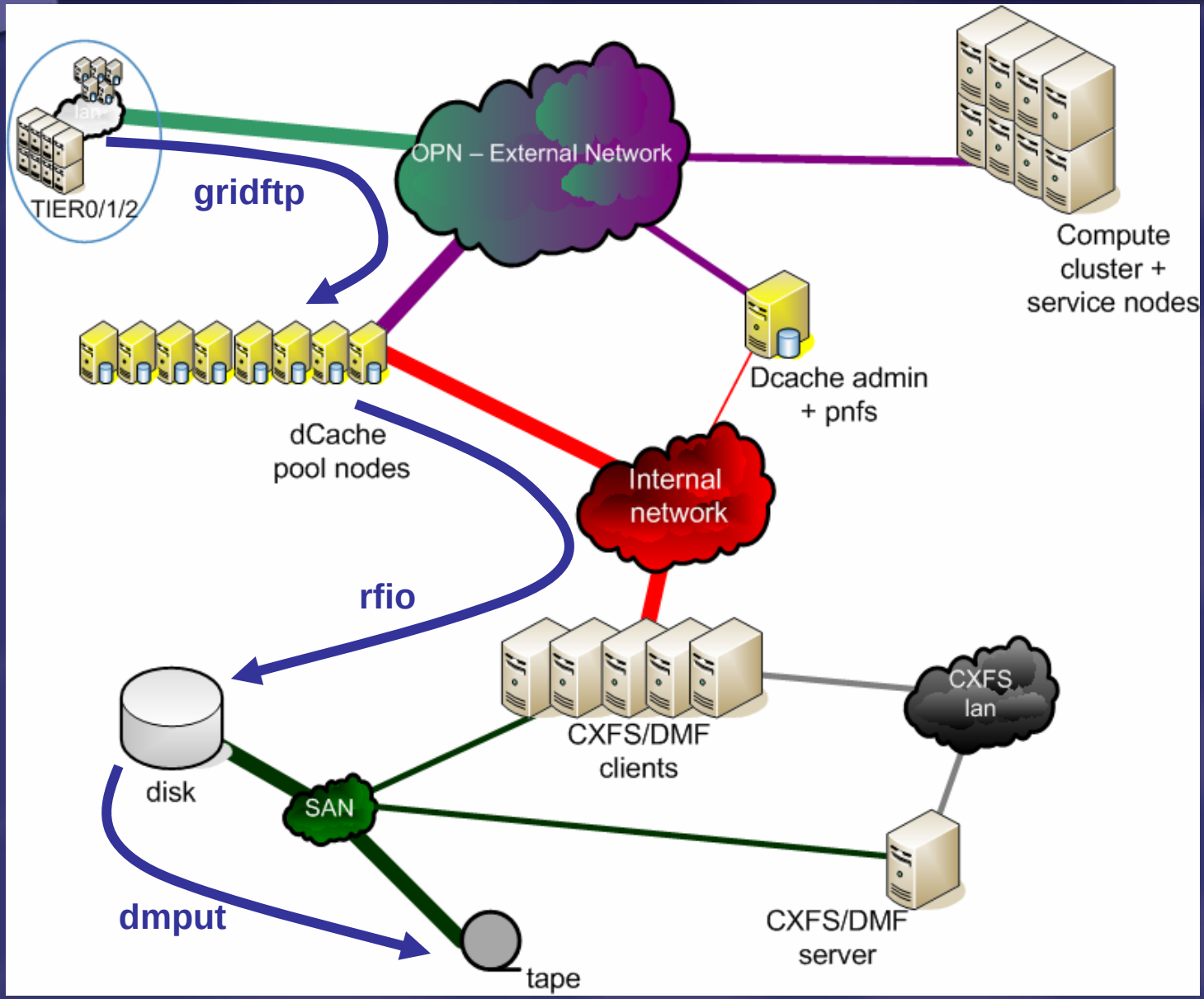
Storage Classes Implementation @ SARA

Ron Trompert

SARA

Storage Infrastructure



Hardware

- **Pool nodes (NAS)**
  - 5x dual Opteron's, 4GB memory, 2x 1GE
    - 1.7TB disk cache, SATA, RAID6, XFS
  - 10x dual opteron, 4GB memory, 2x1GE
    - 6.4TB disk cache, SATA, RAID6, XFS
- **dCache admin node**
  - dual Xeon, 4GB memory, 2x 73GB internal disk, 2x 1GE
- **MSS clients (CXFS/DMF)**
  - 2x dual Xeon, 4GB memory, 2x 73GB internal disk, 2x 1GE, dual HBA FC, 1.6 TB CXFS filesystem (SAN shared filesystem)
- **MSS server (CXFS/DMF)**
  - 4 cpu R16K MIPS, 4GB memory, 12x FC, 4x GE, 2x 36GB internal disk, 1.6 TB CXFS filesystem (SAN shared filesystem), 3x STK 9940B tape drives
- **Network**
  - dedicated 10GE network between CERN – Amsterdam
  - GE internal network between pool nodes and MSS clients nodes
- **Tape libraries**
  - STK SL8500 in Almere
  - STK Powderhorn in Amsterdam

dCache Pools

- Tape pools are read/write/cache pools and disk-only pools are read/write pools
  - No distinction between processing pools and transfer pools
  - Avoid pool2pool copies
- All pool nodes run a gridftp door
- Limited number of pools per node (i.e. ~4)
  - Control the number of movers/load on the pool node

dCache Pools

- Pools dedicated to VOs
  - At least for the ones that move a lot of data
  - Possible to do have dedicated pools for groups/roles within a VO
  - Provides “quota”
  - Easy to get usage data per VO/group/role from dCache without doing “du –ms”

LAN or WAN Pools

- Resources of the SARA/NIKHEF T1 will reside at both SARA and NIKHEF which are separate institutions with their own domain. =>All pools are WAN pools

Data Migration Facility (DMF)

- Migrates data to tape and back
- Automatically migrates data to tape when a filesystem usage exceeds a high water mark. Then data is being written to tape until a second lower threshold in the filesystem usage is reached. Also possible to enforce migration to and from tape (dmput/dmget)

Data Migration Facility (DMF)

- DMF can set different policies related to migration rules per file system and “SELECT_MSP” rules
- SELECT_MSP:
  - Map uids to volume groups which is the same as tape sets
  - Select number of tape copies per uid
  - Uid is not necessarily the uid the VO/group/role is mapped on but the uid used to copy data from the dcache pools to the cxfs/dmf clients

Data Migration Facility (DMF)

- Tape drives can be assigned to volume groups
- Pool of empty tapes is not yet assigned to a volume group are available for everyone
- A tape only contains data belonging to a single volume group
- Performs defragmentation. Partially filled tapes are merged to create as many new empty tapes as possible

Data Migration Facility (DMF)

- It collects write requests and mounts a tape when a sufficient amount of data can be written to tape at a rate close to the optimal rate of 30 MB/s per 9940B drive.
- Collects read requests and minimises the number of necessary tape mounts
  - Read requests should be submitted within a limited period of time.

Storage class implementation

- Currently:
  - T0D1 with non HSM dcache pools at srm://ant1.grid.sara.nl:8443/pnfs/grid.sara.nl/disk/
  - T1D0 with HSM pools at srm://srm.grid.sara.nl:8443/pnfs/grid.sara.nl/data/
- T0<->T1 transitions will not (yet) be supported and T1Dx transitions are only DB operations
  - These transitions not involve the site's implementation
  - T1D0->T1D1 does not involve immediate staging of a file. However, we would still need some advanced way of scheduling staging requests (VO/group/role based) in order to avoid (unintentional) DOS attacks on the tape system

dCache-DMF mapping

- Migration script called by dCache uses the owner of the files to get a uid attached to a specific volume group (tape set) from a hash table
- Up to now we don’t use the dCache storage group info that is contained in the path in PNFS for this but it is possible to do so in the future
- It is also possible here to take the user description of the space token into account when it is passed on to this script
- Migration script copies the data to the cxfs/dmf clients as this uid

File Naming on MSS

- The path in pnfs of a file is reflected in the path on the cxfs clients.
  For example:

  /pnfs/grid.sara.nl/data/atlas/saratest/testfile

  becomes

  /cxfs/TIER1SC/tape_storage/0003/data/atlas/saratest/testfile::00030000000000000068ECE8
- Easy to get information about the owner/type of data and the corresponding pnfsid

Remove files from MSS

- Remote path on CXFS clients is stored in PNFS when file is copied
- Cron jobs collects that info about removed files from /opt/pnfsdb/pnfs/trash/1 and removes the file

dCache prestager

- Every minute a cron scripts collects files queued for restore
  - “rh jobs ls” for all tape pools to see what restore requests are queued and which files are involved
- Scripts issues a dmget for these files