Handling of T1D0 in CCRC'08

- Tier-0 data handling
- Tier-1 data handling
- Experiment data handling
- Reprocessing
- Recalling files from tape
- Tier-0 data handling, cont.
- Tier-1 data handling, cont.
- Experiment data handling, cont.

Proposal for production activities in February 2008

- Executive summary: CASTOR
- Executive summary: dCache
- Executive summary: experiments

Tier-0 data handling

- Central data recording into dedicated service class instances
  - Could be given space tokens if needed
- Operations in parallel:
  1. Write data to tape ASAP such that copy in experiment buffer can go
  2. Distribute data to Tier-1 centers
  3. Make data available to first-pass reconstruction
     - Output also to be written to tape and distributed to Tier-1 centers
- Step 2 may need data to be copied to other service class instance
  - (Better) connected to WAN
  - BringOnline / PrepareToGet would trigger disk-to-disk copy or tape recall
    - Note: the FTS currently expects the data to be available in 3 minutes!
- Step 3 may need data to be copied to yet another instance
  - Better matched to reconstruction program read patterns
  - Reduce interference with steps 2 and 3

Tier-1 data handling (1/2)

- Receive data from Tier-0, other Tier-1 and Tier-2 sites
- Send data to other Tier-1 and Tier-2 sites
- CERN can act as Tier-1, any Tier-1 can act as Tier-2
- Tasks:
  1. Archive raw and reconstructed data received from Tier-0
  2. Make raw data available for second-pass reconstruction
     - Archive output and possibly copy it to other Tier-1 center(s)
  3. Regularly/occasionally recall data sets from tape for reprocessing
     - Archive output and possibly copy it to other Tier-1 center(s)
  4. Serve Tier-2 requests for data sets
  5. Archive Tier-2 Monte Carlo data and certain analysis results

Tier-1 data handling (2/2)

- Disks receiving data from Tier-0 may not be suited for steps 1, 2, 3
  - Need d2d copies
- Data received from Tier-0 must remain pinned for step 2
  - Avoid tape recall
- Steps 2 and 3 need efficient staging of the necessary data
  - Reconstruction jobs should use batch system efficiently
- Large amount of disk space allows tapes to be fully read in one go
  - Allows many jobs to process data in parallel
  - Reduces need for merging small output files
- Dedicated disk space reduces probability of premature garbage collection due to concurrent activities

Experiment data handling

- Small number of production managers
- Large number of unprivileged users, spread over physics groups
- Both categories desire a guaranteed Quality of Service
  - Dedicated disk spaces
  - Certain priorities in various request queues
- QoS handles available in CASTOR and/or dCache:
  - Space tokens
  - Name space
  - User identity/role
  - Client IP address, WAN/LAN flag
- Storage classes
  - Custodial-Nearline (T1D0) managed by system
    - Used for vast majority of the data
  - Custodial-Online (T1D1) managed by VO
  - Replica-Online (T0D1) ditto

Reprocessing

- Large T1D0 data sets will need to be reprocessed
  - Pre-stage from tape, controlled by production managers
  - Disk copies can be garbage-collected right after reprocessing
  - Large amount of disk space desired, ideally dedicated
    - Use BoL/PtG with desired pin time
- Alternative: temporarily change storage class to T1D1
  - T1D1 originally foreseen for data needed online for a long time, even when not used for a while
  - Not implemented

Recalling files from tape

- CASTOR
  - Can recall into an SRM v2.2 space, but original token not remembered
  - Explicit token needed on recall, or system will use other information to determine where the file shall be restored
    - User identity
    - WAN/LAN flag (not yet possible with high-level tools)
    - Retention policy, access latency (ditto)
- dCache
  - Cannot recall into an SRM v2.2 space
  - Sufficient disk space has to be left unassigned to any space token
  - Pool selections allowed through static configurations
    - Pools can be associated with paths

Tier-0 data handling, cont.

- Data can be copied to WAN buffer with given token
  - Files in T1D0 class need to remain pinned for the transfer
  - ChangeSpaceForFiles might be used if no influence on tape writing
- Files might be copied/recalled to default space
  - WAN flag to be provided by FTS
  - Further separation possible e.g. based on user identity
- Reconstruction may also need a separate copy of the data
  - Copy to LAN buffer with token: OK
  - Concurrent use of default space: probably not OK
  - ChangeSpaceForFiles: not OK
    - Moves files from one space to another single space

Tier-1 data handling, cont.

- Considerations for Tier-0 also apply here
  - Very important to separate concurrent activities
- dCache admins prefer separation of read and write buffers
  - Optimize I/O performance
- ChangeSpaceForFiles might be used for (re)processing
  - But it is not available
  - Explicit copying to/from T1D1 (with different file names) does not scale
- Pinning
  - CASTOR: best effort
  - dCache: OK

Experiment data handling, cont.

- Possibly chaotic due to competition between physics groups
  - In particular when no token can be given on BoL/PtG
  - dCache cannot deny read access from pool that should be dedicated to some subset of the VO
  - Production is better controlled
- CCRC'08 should not have a lot of such activity
- CASTOR: user requests should not provide tokens on reads
  - Requests will go to small default pool
  - Avoids recalls going to pool only writable for production managers
  - Avoids copying data that can be served from another service class
- dCache: usage should be as for production

Executive summary: CASTOR

- Configure T1D0 instance sizes as requested by the experiments
- Configure a fraction of T1D0 disk as default space for T1D0
  - Agree sizes with the experiments
  - Take WAN/LAN access needs into account
- Allow for other handles (e.g. user identity) to be used in pool selection when token not specified

Executive summary: dCache

- Keep T1D0 spaces relatively small
  - Use them only as buffers for writing into the storage system
- Keep most of the T1D0 disk space unassigned to any space tokens
  - Can be used for restoring large data sets concurrently
- Possibly configure paths to allow for the selection of specific pools when recalling files from tape
  - Depending on name space layout per experiment

Executive summary: experiments

- Production managers use space tokens for BoL/PtG
  - Allows for a controlled pool selection in CASTOR
  - Allows for gaining experience with "full" usage of tokens
- Normal users do not use space tokens on reads
  - The system decides where files are staged
- Use paths consistently as much as possible
  - Helps optimizing performance of dCache sites