dCache at SARA: Infrastructure, Tuning and Issues

Ron Trompert

GDB 9 september 2008

Contents

- Data access issues and tuning
  - Networkbandwidth
  - Stage from tape
  - Gsidcap
  - Miscellaneous
- Stability and availability measures

Network bandwidth

- The network infrastructure between the SARA compute clusters and the dCache storage is 1 Gbps still.
- Between the SARA dCache storage and the compute clusters at NIKHEF we have 2 Gbps.
- Will be solved asap.

Stage from tape

- We have observed peak rates of 216 MB/s reading files from tape with 6 9940B drives but this very much depends on the file sizes used and activities from other VOs since tape drives are shared. Tuning on both the server as the client side is necessary to improve performance.

Stage from tape: tuning

- File prestaging
  - A dCache stage pool has a maximum number of stage processes that can run simultaneously
  - Additional stage requests are queued
  - A home grown prestage script collects the files that are associated with the queued stage requests and stages them on an intermediate disk cache (on MSS backend, no dCache)
  - When dCache starts handle these queued requests the files are already on disk
  - This gave us a few tens of MB/s extra in read performance

Stage from tape: tuning

- dCache tuning
  - Prevent files from getting into “suspended mode”
    - After a file gets into suspended mode. Staging can only manually be initiated again
    - Increased the number of retries and retry more often (300 retries instead of 3 but try once every 900 seconds)
  - Set the spacecost factor to 1000 for stage pools to get a equal distribution of data over the stage pools
- Tuning MSS backend
  - Load balancing over cxfs MSS backend nodes and put a limit on the load on those hosts.

Stage from tape: tuning

- Tuning on the client side
  - CCRC08 has shown that using big (2+ GB) files makes a big difference. This makes the location of files on tape or on which tape much less critical for the read performance.
    - Right now we have dedicated tape sets for the VOs but is a finer granularity necessary, i.e. dedicated tape sets for different datasets?
  - Know how big the cache stage pools are and how much of it is pinned and do not oversubscribe them to avoid staging the same files over and over again

Gsidcap problem

- A dCache gridftp server can be told to listen to a particular network interface. Lcg-gt then returns a turl associated with the host name on that interface
- This does not work for gsidcap. It cannot (yet) be told to listen a particular interface and the SRM returns turls associated with the hostname on the lowest interface number (eth0). In our storage environment this happens to be an internal management network except for our srm node which has it external network interface on eth0. So therefore we can only run one gsidcap instance on our srm node.

Gsidcap problem

- Since we have only one gsidcap server it needs to serve a lot of logins resulting into memory problems and gsidcap crashing
- This is about to be fixed.

Miscellaneous

- Security on SRM operations
  - With a dteam proxy you can stage ATLAS files on ATLAS pools
- SRM creates directory as root but forget to chown it to the right uid
- The pin lifetime is counted from the moment the srm request was issued, not from the moment the file has been staged
- Reduce p2p copies as much as possible
  - Used to have separate read/write/stage pools. P2p copies proved to be a great performance bottleneck during CCRC08
  - Now TxDx are read/write and stage pools are read/cache
  - This also cures the ONLINE/NEARLINE interpretation issue

Miscellaneous

- Replication of conditions data over all read pools
- Consistency namespace and physical data

Stability and availability measures

- If a problem with a service occurs, a watchdog cron script is put in place that monitors this and restarts the service if necessary
- H/W has at least a dual power supply, RAID system disks and connected to UPS
- Databases are on RAID/SAN and are backed up daily
- Infrastructure monitored by ganglia and nagios