Grid storage - types, constraints and availability

Latchezar Betev

Offline week, April 9, 2008

GRID and CAF user forums

- New initiative, regular (bi-weekly) discussion on GRID and CAF
  - From user perspective – practices, tips, latest news
  - User-suggested topics
  - Immediate expert support, other than by e-mail, Savannah
- First forum on 27 March 2008
  - Telephone conference (no need for specially equipped rooms)
  - 22 participants
  - Positive feedback
- Second forum - today

GRID and CAF user forums (2)

- Forum agenda - flexible
  - The idea is to cover topics suggested by users
  - And to present new development, which is not yet ‘popularly used’
  - Please do not hesitate to propose any topic!
- Announcements
  - alice-project-analysis-task-force
  - alice-off
- Next forum – 24 April 2008

GRID storage

- Basic types – MSS, disk
- What are the differences, constraints
  - In a Grid world, where the user does not know which SE is of what type
  - Which storage to use
- Availability of storage
  - Why do the SEs fail
  - How to figure when the job failures are due to storage
- Production practices

GRID storage types - MSS

- Mass storage System – all data written to this type of storage goes to tape
  - Available only at the large T1 centres
  - Very complex internal structure
- Pros
  - Configured to store very large amounts of data (multi-PB)
  - Still cheaper than disk-only storage
  - Safer, but not by a big margin (see next slide)
- Cons
  - Fast (random) access to data is difficult
  - Disk buffer is much smaller than the tape backend
  - Easy to fall victim to a race condition – multiple users reading different data sample, thus trashing the disk buffer

GRID storage types – MSS (2)

- Why tape (if disk nowadays is cheap and reliable)
  - Strategic decision of all T1s many years ago
  - Investment in tape system is substantial
  - Building of the infrastructure takes a long time
- Current trends
  - Secondary and all tertiary storage functions and utilities, such as disk backup and data archiving, i.e. RAW and ESDs

GRID storage types – MSS (3)

- Storage types
  - dCache – developed at DESY/FNAL
  - CASTOR2 – developed at CERN
- In ALICE
  - RAL, CNAF, CERN – CASTOR2
  - CCIN2P3, FZK, NL-T1, NDGF – dCache
- Both dCache/CASTOR2 implement reading/writing through the xrootd protocol
  - CASTOR2 – plug-in
  - dCache – protocol emulation

GRID storage types – MSS (4)

- ALICE computing model – custodial storage
  - RAW data (@T0 – CERN + one copy @T1s)
  - ESDs/AODs from RAW and MC production (copy from T2s, regional principle)
- From user point of view
  - Reading of ESDs/AODs from MC/RAW data production
  - Writing of ***very important*** files
  - The underlying complexity of the storage is completely hidden by AliEn

Use of MSS in the everyday analysis

- For reading of ESDs – nothing to be done
  - Access typically through collections/tags
  - Automatically taken care of by the AliEn JobOptimizer
  - Users should avoid JDL declarations like

Requirements = member(other.GridPartitions,“Analysis");

  - The above interferes with the JobOptimizer and may prevent the job from running
- For writing
  - ***only*** for copy of important files – JDL, configurations or code, ***never*** for intermediate or even final output of analysis jobs

Use of MSS in the everyday analysis (2)

- Top 5 reasons to avoid writing into MSS-enabled storage
  1. Access to MSS is slow, recall time from tape is rather unpredictable
  2. If your file is not in the disk buffer, you may wait up to a day to get it back
  3. With the exception of very small number of user-specific and unique files, all other results are reproducible
  4. MSS is extremely inefficient for small files (below 1GB)
  5. More and more disk storage is entering production – it is also very reliable, chances that your files will be lost are very small

Use of MSS in the everyday analysis (3)

- Summary of good user practices
  - Use MSS only for backing up of important files, keep the results of analysis on ***disk*** type storage
  - Always use archiving of files. The declaration below will save only one file in the MSS, there is no time penalty while reading

OutputArchive={"root_archive.zip:*.root@<MSS>”};

GRID storage types - Disk

- Disk – all data written to this type of storage stays on disk
  - Available everywhere, T0, T1 and T2 centres
  - Simple internal structure – typically NAS
- Pros
  - Fast data access
  - Price per TB is comparable to tape
  - Very safe, if properly configured RAID, same as tape
  - PB size disk storage can be easily build today
- Cons
  - None really – ideal type of storage

GRID storage types – Disk (2)

- Storage types
  - dCache – developed at DESY/FNAL
  - DPM – developed at CERN
  - xrootd – developed at SLAC and INFN
- In ALICE
  - All T2 computing centres are/should deploy xrootd or xrootd-enabled storage
- Both dCache/DPM implement reading/writing through the xrootd protocol
  - DPM – plug-in
  - dCache – protocol emulation

GRID storage types – Disk (3)

- ALICE computing model – tactical storage
  - MC and RAW data ESDs (T0/T1/T2)
- From user point of view
  - Reading of ESDs/AODs from MC/RAW data production
  - Writing of ***all types*** of files
  - Important files – save 2 replicas (@storage1 and @storage2)

Use of Disk storage in the everyday analysis

- For reading of ESDs – nothing to be done
  - Access typically through collections/tags
  - Automatically taken care of by the AliEn JobOptimizer
  - Users should avoid JDL declarations like

Requirements = member(other.GridPartitions,“Analysis");

  - The above interferes with the JobOptimizer and may prevent the job from running
- For writing - unrestricted
  - Through declarations: file@<SE name>
  - No user quotas yet
  - Easy to change from one SE to another

Use of Disk storage in the everyday analysis (3)

- Summary of good user practices
  - Use disk storage for all kind of output files
  - Report immediately any problems you may encounter (inaccessibility, sluggishness)
  - Preferably use archiving of files. The declaration below will save only one file in the disk storage, there is no time penalty while reading
  - Store 2 copies of your important files at 2 different SEs (maximum safety)

Current SE deployment status

- User-accessible storage

http://aliceinfo.cern.ch/Offline/Activities/Analysis/GRID_status.html

- The local support needs some improvements, however the stability is very reasonable

SE Name	AliEn name	Description	SE Status	Size	Used
1. Subatech - DPM	ALICE::Subatech::DPM	DPM (disk), general use	OK	11.64 TB	0.132 GB
2. SPbSU - DPM	ALICE::SPbSU::DPM	DPM (disk), general use	OK	5.402 TB	1.94 GB
3. Catania - DPM	ALICE::Catania::DPM	DPM (disk), general use	OK	45.63 TB	5.645 TB
4. Bari - dCache	ALICE::Bari::dCache	dCache (disk), general use	OK	4.005 TB	3.651 GB
5. CERN - Castor2	ALICE::CERN::Castor2	Castor2 (MSS), RAW data, ESDs	OK	931.3 TB	475.7 TB
6. CERN - se	ALICE::CERN::se	xrootd (disk), OCDB master, application packages	OK	2 TB	967.9 GB
7. GSI - se	ALICE::GSI::se	xrootd (disk), general use	OK	27.94 TB	20.05 TB
8. Legnaro - dCache	ALICE::Legnaro::dCache	dCache (disk), general use	OK	5.215 TB	930 GB
9. NDGF - dcache	ALICE::NDGF::dcache	dCache (disk), general use	OK	23.28 TB	8.82 TB
10. NIHAM - File	ALICE::NIHAM::File	xrootd (disk), general use	OK	39.12 TB	3.824 TB
11. Prague - Disk	ALICE::Prague::Disk	xrootd (disk), general use	OK	1.267 TB	94.28 GB
12. Torino - DPM	ALICE::Torino::DPM	DPM (disk), general use	OK	16.78 TB	1.015 TB
Total			**12**	**1.088 PB**	**517 TB**

Availability of storage - failures

- Software (predominant, short duration)
  - These gradually go down as the software matures and site experts gain experience in storage maintenance
- Hardware (long duration)
  - Site scheduled/unscheduled downtimes
  - Storage server failures (rare)
  - These will continue to exists on the same level as now, the only continuous data access is replication
    - If sufficient capacity exists

Availability of storage – Job errors

- Two classes of errors
  - AliEn: EIB (Error Input Box), ESV (Error Saving)
    - Obvious
  - ROOT (still AliEn codes): EE (Error Execution), EXP (Expired)
    - A bit more complex – can be also caused by a problems in the code (f.e. infinite loop)
- What to do (as a first step)
  - Check SE elements status
  - Do not attempt to read data not staged on disk (check ‘staged’ status)

Production practices

- For efficient analysis the ESDs + friends should be on disk
- So far, the predominantly used storage was MSS@CERN
  - This is quickly changing in view of the rapid deployment of disk storage at T2s
- The output from the presently running productions (LHC08t, LHC08p LHC08u) is saved at T2 disk storage + copy @T1
- All past productions are staged on request on MSS and replicated to T2 disk storage

Summary

- The storage availability and stability is still the Grid's weak point
  - The progress in the past 6 months is substantial – from 2 SEs to more than 15 used in production
  - The stability of storage is also improving rapidly
- New disk-based storage (at T2 sites) allows for more efficient data analysis
  - The primary copy of the output files of recent productions is stored at T2s (disk)
  - Old productions are replicated to T2s
- User Grid code should be modified to take advantage of the new storages
- Please report problems with storage immediately!