ATLAS Note draft
10 November 2006 (V1.2)

ATLAS Data Transfer Functional Test (October 2006)

F.Brochu, J.Chudoba, C.Condurache, W.Deng, X.Espinal, S.Jezequel, H.Ito, J.Kennedy, A.Klimentov, G.Negri, P.Nevski, C.Serfon, J.Shih, R.Walker, X.Zhao, S.Zhou

ATL-SOFT-PUB-2006-009
06 December 2006

Data transfer functional test was conducted from October 22 to October 30, 2006 by DDM operations team.
ATLAS DDM software components were used to transmit data, control and monitor data movement. The main purpose of the test was to check system functionality during data transfer from CERN to ATLAS Tier-1s and data transfer within Tier-1 clouds. Two additional cases : large files (files with size larger than 2 GB) and data transfer between Regional Centers have been tested.
Centers participated in the test are listed in Table 1.

Tier-1	Test Coordinator	Tier-2, Tier-3 sites
ASGC	Jason Shih	IPAS, Uni Melbourne
BNL	Hironori Ito	AGLT2, BU, MWT2_IU, OU, SLAC, UC_VOB, UTA_SWT2
CNAF	Guido Negri	LNF, MILANO, NAPOLI, ROMA1
FZK	John Kennedy	CSCS, CYF, DESY-ZN, DESY-HH, FZU, WUP
LYON	Stephane Jezequel	BEIIJING, CPPM, LAPP, LPC, LPHNE, SACLAY, TOKYO
PIC	Xavier Espinal	IFAE, IFIC, UAM
RAL	Frederic Brochu	CAM, GLASGOW, LANCS, MANC, QMUL
SARA	Jiri Chudoba	ITEP,SINP
TRIUMF	Rod Walker	SFU,TORONTO,Uni Montreal,VICTORIA,ALBERTA

Table 1: ATLAS centers participated in data transfer functional test

The test scope

- Data transfer from CERN to Tiers for datasets with average file size 400 MB. This step simulated the flow of ATLAS data from Tier-0 (CERN) to the sites.
  - Step 1 : Data transfer from CERN to ATLAS Tiers
    - a. data tansfer from CERN to Tier-1
    - b. data transfer from Tier-1 to Tier-2s
  - Step 2: Data transfer from Tier-2s to Tier-1. To simulate the Monte-Carlo production data flow.
- Large file transfer. Transfer data from CERN to Tier-1s and then from Tier-1s to the selected Tier-2s.
  - 1 reference dataset with 3.1 GB files
  - 9 datasets (one per cloud) with 4.5 GB files
- Data transfer between regional centers
  - Tier-1/Tier-1 data transfer
  - ”foreign”Tier-2/Tier-1 data transfer

1 Data transfer from CERN to ATLAS Tiers

The data transfer from CERN to Tiers was the most successful step of the test. 72 datasets (2000 files) have been sent from CERN to Tier-1s and then to Tier-2s. One reference (aka *global*) dataset and one dataset per each Tier-2 were generated using ATLAS MC production. The subscription for datasets was done simultaneosly from all Tier-1s to CERN. The average failure rate was less than 5% for all sites but TRIUMF. Most of the problems for other sites have been fixed without test interruption. The failure rate for TRIUMF was more than 50%, and the reason is not completely understood. Despite different actions on the site and by DDM developers have been taken, but we didn’t succeed to complete even the first step of the test. The problem was solved by deleting local DQ2 database and DB tables initialization.

The data transfer within clouds was done in two directions. Data have been sent from Tier-1 to Tier-2s, deleted from Tier-1 and then copied back. There were different failures in different clouds mostly related to FTS configuration. The summary for all clouds is following.

2 Large file transfer

The large files (files with size more than 2 GB) transfer was tested. 10 datasets have been prepared. The reference dataset from Peter van Gemmeren with 3.1 GB files (all files with the same internal GUID!) and 9 datasets (1 per Tier-1) with 4.5 GB files. The tested input/output SE combinations are listed in the Table 3. The I/O throughput for 4.5 GB files between CERN and BNL was 29 MB/s [1)]

[1)] I/O throughput is calculated as number of transferred GB divided by elapsed time, where elapsed time is an interval from data transfer initiation to data transfer completion

Tier-1	Status	Comments
ASGC	done	failed for AU-UNIMELB. FTS (?) channel configuration
BNL	done	
CNAF	done	
FZK	not completed	problems after dCache upgrade at FZK
LYON	done	The FTS server in LYON wasn't able to manage 6 FTS channels simultaneously.
PIC	done	high failure rate for UAM. probably network glitch and then DQ2 SW problem
RAL	done	FTS authorization problem for Edinburgh. fixed (DC)
SARA	done	failure rate slightly higher than for other sites
TRIUMF	failed	50% failure rate between CERN and TRIUMF 100% data transfer failed from TRIUMF to ALBERTA and SFU

Table 2: Data transfer from CERN to ATLAS centers

Sites (SRC-DEST)	Storage Elements	Status
CERN-ASGC	CASTOR-CASTOR	done
CERN-BNL-AGLT2	CASTOR-dCache-RAID5(local)	done
CERN-BNL-BU	CASTOR-dCache-DPM	failed from BNL to BU
CERN-BNL-SLAC	CASTOR-dCache-NFS	done
CERN-CNAF	CASTOR-CASTOR	done
CERN-FZK	CASTOR-dCache	done
CERN-LYON-LAPP	CASTOR-dCache-DPM	failed from LYON to LAPP (*)
CERN-PIC-IFIC	CASTOR-dCache-CASTOR	failed from PIC to IFIC
CERN-PIC-IFAE	CASTOR-dCache-UnixFSSrm (**)	done.
CERN-RAL	CASTOR-dCache	done
CERN-SARA	CASTOR-dCache	done
CERN-TRIUMF	CASTOR-dCache	done

Table 3: Large files transfer

(*) transfer was blocked probably by the FTS server in Lyon
(**) UnixFSSrm = SRM dCache + gridftp classic + posix file acces[2)]

2.1 Problems and uncertainties with transferring and storing files larger than 2 GB

.

- CASTOR : Large files support requires CASTOR SRM-2.2.9-4 and CASTOR client release 2.1.1
- dCache :

[2)] https://srm.fnal.gov/twiki/bin/view/SrmProject/UnixFSSrm

  - PNFS can only represent a data file's size accurately up to (2G-1)B; beyond that, the file size is shown as 1. Enstore knows, stores and uses the real file size.
  - 2+GB file copying from dCache to the local file system using *dccp* fails
  - 2+GB file copying from dCache to the local file system using *globus-url-copy* and *srmcp* works correctly

- DPM. Data transfer failed to two DPM SE (Boston University and LAPP). We cannot confirm that both cases are related to the SE flavour. The test need to be repeated.

3 Data transfer between Regional Centers

ATLAS centers associations schema was used to select pairs for data transfer between different regional centers.

Data transfer between regional centers

Monitoring

There were problems with both monitoring packages which significantly complicated the test. ARDA monitoring was slow, unstable and unreliable. DDM monitoring was not always up-to-date. That is why most of the data transfer checks have been done on sites using logs, and local files catalogs.

DQ2 client tools

DQ2 client tools and special scripts (like DQ2 cleanup procedure) were used for datasets subscription and control. All tools worked correctly.

Summary and Conclusions

The data transfer robustness is improved since the previous test in September. The problems on sites (FZK and TRIUMF) related to the site and ATLAS DDM SW persisted for the whole test duration. The data transfer between regional centers needs more testing. For ASGC cloud the FTS channel setting with Uni.of Melbourne must be checked.
ATLAS DDM system can cope successfully with large files transfer.
Crash of central DDM DB server happened several times during test week. The problem is under investigation. Probably crashes are related to one of DQ2 client commands (file lookup).

DDM Operations Team in October test

F.Brochu, J.Chudoba, C.Condurache, W.Deng, X.Espinal, S.Jezequel, H.Ito, J.Kennedy, A.Klimentov, E.Lyublev, P.McGuigan, S.Mckee, G.Negri, P.Nevski, C.Serfon, D.Schragger, H.Severini, J.Shih, R.Walker, W. Yang, S.Youseff, X.Zhao, S.Zhou

DDM developers support : M.Branco, D.Cameron, P.Salgado

Related Documents

- X.Espinal. Tier-1 and Tier-2 Functional test at the Spanish cloud. IFAE/ATLAS/LCG/06-01, Oct 31st, 2006. ATL-COM-SOFT-2006-015.
- M.Branco . Comments from DDM developerss on the problems with October Data Transfer Functional Test. Nov 3rd, 2006 Private communication (attached)

Attachement: Miguel Branco. Note on the problems observed during October Data Transfer Functional Test

The first comment is that problems with TRIUMF were not a problem with the site, but with DQ2, its monitoring service and other unexpected data flows (not part of the October functional test). [more details below]

We suffered instabilities in our atldq02 machine (the monitoring server for DDM monitoring), which is a single machine, slightly better than a desktop machine, that served as our monitoring server, receiving monitoring requests from all DQ2 services worldwide.

We did not immediately find the problem: in fact the machine was not working properly for several days already. As the transfer tests started we then finally saw the problems and fixed them but most sites had accumulated a very big backlog of monitoring information.

In some VO BOXes, this backlog was very slow to recover. After fixes on the DQ2 local database (indexing MySQL) this was recovered. The problem was particularly difficult to discover as it did not happen in all sites (in some high-end VO BOXes, the lack of DQ2 local database indexes was not noticeable which made debugging more confusing)

We also had to change the monitoring information for the case where data would be at the site, removed, then copied back: we simply never had this case before so the way records were stored in our monitoring database did not support it.

Note that we did not expect to do any development on our monitoring service but move to ARDA service - which we could not do because ARDA was facing scalability problems and decided to go back to development of an Oracle backend, as opposed to a Postgres backend as they had running.

Also note that in some of the sites - notably TRIUMF - other transfers were occurring, T2-T2 (from Cracow to a TRIUMF T2) and these were failing because the source site was DPM (does not support srmCopy). The FTS channels for these transfers were not setup on DQ2, therefore the transfers were failing but definitely clogging the system. So during the tests we had to reconfigure DQ2 FTS channels to support all these data flows, very often kill hanged SRM processes on the VO BOX, even though they were not part of the functional test: but these were indeed affecting the performance

of the functional test, by affecting the site services.

We would say monitoring was the most important problem in the exercise - due to a combination of server problems, no development plans for our monitoring service, missing functionality and unexpected scalability problems on the ARDA side. Additionally, problems transferring data to some of the Tier-2s and the existence of other transfer flows not part of the test, affected the overall functionality.

We realize we may need to rework our monitoring service or move quickly to an ARDA service in production. We will need to improve our replica selection choice for alternative routes: being worked on and expected to be complete soon. We have already added FTS wildcard channels serving Tier-2s to always use FTS. We will be improving our retrial and subscription 'degradation' policy to prevent problems with too many retrials. We would like to point out the case of FZK, a Tier-1 unavailable for over a week, with DQ2 having to compensate with a constant retrial (with a gentle backoff). We have fixed local database issues when the monitoring information was unable to be dispatched to central services and would be retained in the local site database.

Overall, the exercise was extremely useful, in particular the daily reports. At this moment we believe it is not useful to schedule another round of exercises until DDM team has time to address some of the issues seen during this exercise (which should not happen in the next 4 weeks at least, with the DDM review and arrival of new developers). After these issues have been tackled, it would be very useful to rerun the test!