

My (KB) notes taken during the April 24 2008 ATLAS T0/1/2 Jamboree

(April 28, Version 1)

(May 6, Version 2, Changed ATLASGRP to ATLASGROUP everywhere)

These notes are a reflection on what was discussed during the session. They don't try to summarize what is in the slides that were presented. The slides are available from the agenda page: <http://indico.cern.ch/conferenceDisplay.py?confId=22138>

1. Test plan for May (Kors)

Week 1: Functional Test at nominal rate for 3 days

Just 1 copy of the ESD to T1's according to share.

Week 2: T1 – T1 tests using the ESD.

If all goes right afterwards each T1 should have the full ESD copy.

Week 3: Throughput Test

Run 3 days at 100%, 150% and 200% of the rates by oversubscription.

Week 4: Contingency

We need the T1-T1 channel to distribute AOD and ESD output from the re-processing to the other T1's. This data must also be archived to tape at the T1 where it is produced. Therefore this must be in the T1D1 type of storage.

However, since we realized that the transition from one storage class to another is not implemented we decided not to use this storage type. This transition is needed (from T1D1 to T1D0) when files must be deleted from disk but kept on tape. We decided instead to implement it as a T1D0 pool and use pinning to avoid the data from being garbage collected.

So T1 sites are requested to remove any T1D1 disk space if it existed already.

[Action on: all T1 managers](#)

And T1 sites are requested to make 2 TB of disk space available of T1D0 with a pinning lifetime of 2 months.

[Action on: All T1 managers](#)

The instructions how to set up such a pool are available and links will be distributed and put in the appropriate wiki's.

[Action on: Flavia Donno](#)

The slides and the wiki have to be brought up to date to reflect this decision.

[Action on: Kors Bos](#)

2. Resources (Jim)

We can make a reasonable estimate now on how much storage is needed for early data taking but we don't have this yet per storage class: how much is needed in the MC pool compared to what is needed for the detector data or what is needed for the re-processing.

A table by data-type needs to be made reflecting all storage we need.

[Action on: Jim Shank](#)

3. New Storage Set-up at CERN (Simone)

The t0 is now completely on SRMv2 and the v1 service has been discontinued.

All t0 pools should be protected against normal users accessing any of the files. The atldata pool contains detector data and the atlprod simulation data. Both can only be written by production managers but read by all ATLAS users. The atlcal pool is for calibration and monitoring data and is managed by the involved groups. The default pool is the only other ATLAS pool with a tape back-end. It is meant for all user files at CERN but all ATLAS people globally have access to it.

Sites have to make sure only production managers can write in the data- and Monte Carlo pools. Users should never be able to read files from tape.

[Action on: T1 site managers](#)

4. SRMv2 + Space Token Migration (Stephane)

ATLAS now assumes all sites T1 and T2's use SRMv2. For CCRC in May we will still allow T2 sites that have not made the conversion but we consider this an exception. The central operations team will need to test that all storage spaces have been correctly set up.

[Action on: some last T2 sites to convert to srmv2](#)

[Action on: central operations to test all srmv2 spaces](#)

We now need space for physics groups to write their output (primary DPD) data. Contrary to what we said in the last Jamboree we now only foresee one storage space for all groups. The space token is ATLASGROUP and the storage type is disk only. It is expected that it will be little used in May (almost no physics data) so sites are requested to install not more than 2 TB.

[Action on: T1 and T2 sites to prepare storage for ATLASGROUP](#)

The distribution of AOD's in the T2's is a matter of the cloud. In clouds with a small number of T2's they may choose not to have another full copy of the AOD's in the T2's.

5. End-User storage space

There was a long discussion on disk space for end users. The original model was to provide "nearby" scratch space for temporary storage. However we

have no or very few tools to manage such a scratch space. If we would have local policies everywhere for such space the success of a job could become very depending on where the executable were run and the storage solution chosen.

We therefore decided to go for the option of reserved storage space for each user within its cloud. Where clouds correspond to countries (the majority of cases) this is a manageable solution. For example all French users can save the output of their job in the site where it ran and if there is a problem with that storage element they can fail over to one of the other French T2's or to the storage in Lyon. For the T2 sites that don't have a T1 in their country we need to look at another workable solution one by one.

This is ATLAS managed space although we rely strongly local management for the disk space and the quota. However it is part of the pledges and the files should all be registered in the catalogs. Whenever a user decides to work on a local copy outside the catalog it becomes T3.

Since it is not managed centrally, it is left to the sites to decide on the implementation. Most likely sites will make it available as a managed disk-only. In the original calculations 1.5 TB/user globally were foreseen in the computing model so with 800 users and 50 sites this amounts to something between 10 and 20 TB per site depending on the T2 and the cloud. Because we don't have 800 users yet, it is not expected that this amount will be needed in May.

[Action on: T1 and T2 sites to set up this end user storage space](#)

6. Data Deletion (Vincent)

Data deletion (remove files and replica's and update catalog) is a service now and the goal of 1 Hz at the Tier-1's has been obtained in almost all cases now. To not overload the SRM at the sites the service throttles the deletion process. It was asked if the same service could be provided for moving data (copy + delete).

7. Data Pre-Staging (Birger)

Pre-staging with SRM works on all sites. Performance on srmLs and getStatusOfBringOnline is better on smaller (10 rather than 100) number of files. We will preferably use srmLs because it is independent from the bringOnline request.

8. Re-Processing (Rod)

All small steps in re-processing have been shown to work almost everywhere, we now have to scale it up and turn it into a service. We will use the M5 data: 10 data sets with 250 files each and a total of 5 TB and 1 big data set with 5000 files of a total of 10 TB.

Action on central operations team: if needed, (re-)subscribe data sets to tape in T1's

Each job will need simultaneous access to many (34) Conditions Data files. This was a problem in many sites. Try now to make them into 1 tar ball and send that with the job to the worker node. From the local disk there is not that access problem. If the tar ball is the solution this needs to be provided and distributed centrally.

An intermediate solution exist for pre-staging using the PandaMover but we would like a DDM service using SRM commands.

Action for DDM team: develop a pre-staging service

Files need to be pre-staged in the T1D0 disk pool and stay there sufficiently long for a pilot to be able to pull in a job when a file is available. After the file has been processed it can be purged from the pool. The pool should be big to make pre-staging efficient but not this time so we forcefully learn to throttle the process.

Action on T1 sites: if not done already, make (2 TB) T1D0 disk space available to pre-stage RAW data files from tape.

The ESD and AOD (not for M5) data should be saved on tape, remain on disk for some months and be replicated to other T1's. Because of the decision this morning we will now request T1D0 space with pinning for these data. When newer versions of those files have become available after subsequent re-processing efforts, older versions may be un-pinned and garbage collected from the disk.

Action on T1 sites: make (2 TB) T1D0 disk space (with pinning) available for ESD and AOD from re-processing (see also point 1.)

A separate test is needed to test the use of the database for the conditions data. For this the FDR-1 data will be use for re-processing.

Action on Sacha: to make a detailed plan for such a test

9. Communication and Shifts

The alarm mailing list has been set up (but not tested) in Amsterdam and at CERN (atlas-grid-alarm@cern.ch). Each site is requested to provide a similar service by which an email to this lists triggers an action to resolve the problem. Each site may have its own implementation but the result should be the same.

The MoU describes that we may expect a reaction to such mail within hours during off-working time and a resolution of the problem within hours again depending on the problem. The way of a site to acknowledge reception of the message and further communication is through the computing operations mailing list: atlas-project-adc-operations@cern.ch

Currently only 4 people at the T0 can write to this mailing list. If we run shifts globally this needs to be extended such that there is always such a person

available in the day light time zone where the shifters are. It should remain restrictive because of the sequence of actions that may be triggered.

[Action on Xavier: to propose a new list of names for this list](#)