ATLAS Analysis ST 82
FR Cloud Feedback

21-Jan-09

F.Hernandez, I.Ueda, S.Jezequel, F.Chollet, E.Lançon

To: Daniel van der Ster and Johannes Elmsheuser

FR Cloud ST 82 : http://gangarobot.cern.ch/st/test_82/

Summary of problems tracked are available there:
http://lcg.in2p3.fr/wiki/index.php/Atlas:Analysis_Challenge-STsummary

Introduction

This short report aims to provide input to ATLAS regarding the analysis tests performed on November and December 2008 involving several LCG-France sites. It is important to remember that most of the sites engaged in this exercise are also supporting other LHC experiments. As a consequence, site administrators may not be informed of all the details of ATLAS computing activities as administrators of ATLAS dedicated sites could be.

Targeted metrics and monitoring

The fact that Site Stress Tests uses a test framework in conjunction with a tool parsing the job outputs and publishing the results is very helpful to follow the targeted metrics : Event Rate (evt/s), Success rate (success/failure rate), Job timings and CPU utilization.

Sites are very much concerned about optimizing the data analysis on their site and are ready to look after the infrastructure to see how it performs for ATLAS analysis activities. To be successful in this matter, the optimization process should be a coordinated effort both on the ATLAS software side and on the site infrastructure side.

Analysis job description, sites concerns

In order to understand the observed error messages, experts from sites need to understand “what the analysis job does”.

It would be extremely useful to have a description of the various phases of a typical ATLAS analysis job used for the ST (i.e. Athena Software Setup, Prepare Inputs, Athena Running, Output Storage) with the “running conditions” (namely the number and size of input and output files, input datasets, protocols used for accessing those data, ...).

Sites have expressed concerns about the stress over the software area induced by the Athena software set-up phase. Our understanding is that any single ATLAS analysis job performs the same setup phase. Running this phase simultaneously by several hundred jobs puts a high load on the ATLAS software area. Is this intended to evolve ?

Test conditions, plans for 2009

The FR-Cloud is in favour of running these tests in a controlled manner on a regular basis. We are currently thinking of the proper way to do so but we have already some remarks and suggestions.

We suggest that the same input datasets be used on all sites so that test results (especially job timings) can be compared significantly. For this purpose, we have started the replication of more AODs within the FR cloud.

Provided the test conditions are known and can be reproduced, it might be useful for a site (or for at least a cloud) to perform such tests on demand (autonomously, may be through a web request form). Although it requires the active participation of the site administrators our feeling is that it should be much more efficient to address site-specific problems. Is this possible and compatible with ATLAS ST plans?

Input files access:

- **Changes in stage-in script**

Since the stage-in script and other components of the FT software are still evolving, it would be necessary to rerun ST tests under controlled conditions regarding sites and ST-software changes. Is this foreseen?

- **Use of lcg-gt:**

We have observed jobs being blocked, trying to open a file with a foul TURL:
`rfio:/dpm/in2p3.fr/home/atlas/atlasmcdisk/…` (no dpm pool specified)
in place of `rfio://marjoe.in2p3.fr//baie_atlas3/atlas/…`

One can suspect the lcg-gt command did not return any value so the stage-in script kept some default value. Ganga may adapt the fail-over mode for this case. Still it is not clear if this was caused by a hick-up of the SRM (no special load observed by site on the DPM headnode) or by a problem accessing remote Top BDIIs. Two DPM sites in the cloud have observed this problem; both of them do not have Top level BDII.
So is it possible to use lcg-gt with the –b, --nobdii option to avoid the BDII query?

- **rfio tuning :** tuning of the RFIO read-ahead buffer size on the worker nodes should be probably considered by some sites (not implemented by now)

Multi / co-located sites (GRIF and CC-IN2P3 cases):

Starting from DQ2 site names (GRIF-LAL_MCDISK, GRIF-SACLAY_MCDISK, GRIF-LPNHE_MCDISK), Ganga framework uses *TiersofATLAS to identify* 3 GRIF storage elements at LAL, SACLAY, LPNHE. Then, it queries the information system to found out which computing elements are close to each one of these storage elements.

Given that GRIF is a distributed site, the 4 computing elements at LAL, SACLAY, LLR, IPNO are currently configured to have equal access to GRIF-LAL and GRIF-SACLAY storage elements via the private optical network (GRIFOPN) being set-up between all those sites.

We observed during test 82 that jobs generated for GRIF-SACLAY were executed on the computing element at LAL with inputs taken from SACLAY storage element and outputs written on LAL storage element. A good test the 5 Gb/s link between SACLAY and LAL has been performed.

For now, site configuration seems to be transparent to ATLAS but it is important to take into account the special case of GRIF. For mid-term work and GRIF optimization, GRIF people might be interested to see if performances are affected or not by their strategy and to perform stress tests CE by CE. Is this possible?

Co-location of a tier-1 and a tier-2 at CC-IN2P3 has required the blacklisting of the tier-1 computing elements in the Ganga framework during these tests. We suggest considering a mechanism for creating a JDL specifically requesting the WMS to submit the jobs to sites publishing the tag" VO-atlas-tier_T2".

We would like to thank you for your efforts to set up those tests and your appreciated support.