LCG-France Tier-1 @ Analysis Facility

Summary of Meeting with ALICE Representatives

March 4th 2009 – 14h00-17h00

Attendees :

- ALICE : Federico Carminati, Latchezar Betev, Fabrizio Furano, Fons Rademakers, Yves Schutz
- CC-IN2P3 : Jonathan Schaeffer, Jean-Yves Nief, Yvan Calas, Pierre-Emmanuel Brinette, Benoit Delaunay, Pierre Girard, Ghita Rahal, David Bouvet, Dominique Boutigny, Fabio Hernandez

Chairman : Fabio Hernandez (and Ghita Rahal for the AOB)

Secretary : Fabio Hernandez

Location : CC-IN2P3, Room 202

Agenda : http://indico.in2p3.fr/conferenceDisplay.py?confId=1773

Starting time: 14h30

1. Introduction

The purpose of this meeting was twofold:

- Explore the possibilities of testing a xrootd-based storage elements for Alice at LCG-France tier-1 operated by CC-IN2P3
- Discuss the plans of CC-IN2P3 regarding the prototyping of a PROOF-based interactive analysis infrastructure for the LHC experiments

This document presents the summary of what was discussed and agreed regarding the 2 topics above.

2. Xrootd-based storage elements for ALICE

The storage element for ALICE at CC-IN2P3 is currently managed by dCache/SRM. The file transfers from the tier-0 to CC-IN2P3 are scheduled by the FTS service (at CERN) using the SRM interface to negotiate the transfer protocol (gridFTP in this case). Data transferred to CC-IN2P3 are permanently stored on tape (HPSS). dCache is configured to allow ALICE jobs running at CC-IN2P3 to access those files through the xrootd protocol.

After discussion, we agreed to deploy a test setup dedicated to ALICE based on xrootd servers. This testing environment will be made up of 3 main components, to address several needs:

- **Storage component for importing data from tier-0**: the transfer of RAW data from tier-0 (CERN) to tier-1 (CC-IN2P3) will be handled by SRM/dCache and scheduled by FTS, currently running at CERN. The data will be managed by dCache, which will store it permanently in HPSS, as needed. This component is already in place and is configured to reach the availability targets set by the WLCG project for a tier-1. It was agreed that no changes are to be implemented for this component.

- **Tape-backed storage component for serving local and remote ALICE jobs**: a dedicated xrootd-managed pool of disk servers will be setup by CC-IN2P3 for serving

read-write requests by local ALICE jobs (i.e. the ones running at CC-IN2P3) and remote ALICE jobs (i.e. the ones running at other grid sites). This pool will then be accessible through both the local and the wide area networks.

ALIEN, the ALICE-specific middleware, features an authorization mechanism to files in this component. This mechanism also plays a regulation role for avoiding overloading this storage component by potential users external to the ALIEN framework. The mechanism is based on the distribution of "access tokens" to the ALICE individuals authorized to access the data. ALIEN logs all the access requests to the files in this storage component and the logged information is available to CC-IN2P3 on request.

Should the requested file not be present on one of the xrootd-managed disk servers, it will be staged in from either dCache (for data on disk) or HPSS (for data on tape), depending on its physical location.

This storage component will also be configured for writing into it. ALICE will use this functionality for transferring data from other grid sites (other tier-1s and tier-2s) to be stored at CC-IN2P3. Files will be written into this component by using the 'xrdcp' command and will be permanently stored in HPSS. The location of those files within the HPSS namespace is a decision taken at the level of the xrootd migrator agent.

- **Disk-only storage component for serving local and remote ALICE jobs**: a dedicated xrootd-managed pool of disk servers will be setup by CC-IN2P3 for serving read-write requests by local and remote ALICE jobs. As for the tape-backed pool, this pool will be accessible through both the local and the wide area networks.

  The management of the available space of this pool and also of the contents of it is under ALICE responsibility. In case of exhaustion of the limited disk storage served by this component, ALICE will perform the necessary garbage collection to free up some space, if needed.

CC-IN2P3 wishes to continue using xrootd on SunOS (using ZFS). Fabrizio will verify the availability of the appropriate version of the xrootd software, the one including the authorization mechanism needed by ALIEN. Jean-Yves Nief is the contact person for xrootd matters at CC-IN2P3.

3. PROOF-based interactive analysis prototype infrastructure at CC-IN3P3

Fabio presented the foreseen configuration of a PROOF-based analysis farm at CC-IN2P3 (see slides attached to the agenda). The prototype farm would be composed of around 100 cores and several dedicated file servers exposing the xrootd protocol.

After the discussion on the pros and cons of the envisaged configuration, here is the summary of Fons' recommendations:

- ALICE has shown that a 30 CPU cores farm can serve 10 typical ALICE interactive users, with an aggregated CPU load of 75%. As an example, for some data mining purpose comparable to the analysis activities in HEP, Google target is to keep the farm below 60% CPU load to give its users the perception of interactivity.
- In the case of the data be served by local disk (i.e. internal to PROOF worker node), 1 disk spin can feed 2 CPU cores, and 1 SSD can feed 8 CPU cores. In addition, a configuration in RAID-0 (i.e. mirroring) is recommended for performance.
- Regarding the connectivity of the worker nodes, PROOF is aware of the network topology of the worker nodes. It is recommended to interconnect worker nodes in the same rack with a high-speed switch and interconnect PROOF racks with 10 Gbps links.

The issue of importing users' data into the file servers of the analysis farm was discussed. For the first iterations of the prototype, users will "manually" copy (via the xrdcp command) the desired files into the xrootd pools of the analysis farm.

4. AOB

- Deployment at CC-IN2P3 of a CREAM-based computing element
  CC-IN2P3 is working on the integration of the local batch system (BQS) to the CREAM framework. The necessary components to develop the BQS-specific plugin are not yet provided by CREAM. Other batch systems, such as PBS or LSF, have been interfaced to CREAM by parsing log files, which CC-IN2P3 would like to avoid. The usage of the available BQS APIs is considered the correct way for doing this work for the long term. Close contact is kept between the CC-IN2P3 expert and the CREAM developers team.
- Hardware upgrade of the ALICE VO-box at CC-IN2P3
  A hardware upgrade is planned for all the VO boxes at CC-IN2P3, including ALICE's one. This is to solve a problem observed with the hardware currently being used for those machines, all from the same vendor. This process is to be finished by the end of March.
- Jobs ALICE failing
  Hundreds of ALICE jobs were repeatedly detected stalled, apparently having problems contacting the local VO-Box. Latchezar will look at this issue.
  [Post-meeting information: on March 5th, Latchezar confirms that the problem has been identified and solved in the next version of ALIEN, which was installed at CC-IN2P3 on the same day at 11am.]

5. Next meeting

Not scheduled.

Ending time: 18h10