Grid Interoperability and Massive Data Transfer – IGTMD Project

Minutes of the meeting held at CC-IN2P3 on February 27, 2006

Attendees: Philippe d'Anfray (RENATER), Jérôme Bernier (CC-IN2P3) - Dominique Boutigny (CC-IN2P3), Fabio Hernandez (CC-IN2P3), Franck Simon (RENATER), Dany Vandromme (RENATER), PascaleVicat-Blanc-Primet (ENS / LIP).

This is the first meeting of the IGTMD project after the final acceptation from the ANR. IGTMD is funded for 3 years at the level of 360 k€.
RENATER and ENS/LIP have received their money for the first year. At the time of the meeting, CC-IN2P3 had not received anything yet. The credit notification letter has been received on March 6.

Manpower

It is now possible to hire personnel for the project.

- RENATER: Philippe d'Anfray is the engineer in charge of IGTMD
- ENS/LIP : Pascale Vicat-Blanc-Primet will start looking for a network engineer
- CC-IN2P3: A 2 year Post-Doc in the network field will be hired. Pascale will take care to write a job description and to distribute it widely (see Pos-Doc description at the end). The Post-Doc will be located at CC-IN2P3 but will be driven by Pascale. Argonne is a possible good place to search for this Post-Doc.

Material

Concerning the hardware to be bought for IGTMD:

- Pascale has been looking for 10 Gbit/s programmable network cards. PCI-X cards from CHELSIO seem to be very interesting. The price is ~3-4 k€. She is also considering MYRICOM Myrinet / Ethernet cards. The plan is to buy 2 cards this year.
- Jérôme is planning to buy 3 or 4 servers this year. The idea is to buy 1 Gbit/s configurations this year and to upgrade to 10 Gbit/s next year. Pascale and Jérôme will define the exact configuration for these servers and Jérôme will place orders.

Network

The network infrastructure is obviously very important for this project. Several possibilities exist to transfer data between CC-IN2P3 and FNAL. The simplest one is to use the general IP network through RENATER and GEANT. Given the current usage of the network it seems possible to get a few Gbit/s bandwidth, provided that CC-IN2P3 gets a 10 Gbit/s dedicated connection to REANTER / GEANT.
In a second step it is possible to get a 2×1 Gbit/s dedicated between CC-IN2P3 and New-York, this link could possibly be upgraded to 10 Gbit/s next year.

We should establish a contact with the relevant people in the US (ESNET) in order to understand what is the situation between New-York and FNAL, and what can possibly be done to extend a dedicated network up to FNAL.

At CC-IN2P3 Laurent Caillat will take an active part to the network experimentation.

Interoperability

Two persons from IN2P3 will work on interoperability. Hélène Cordier will work on the interoperability from the point of view of inter-grid operation and Sylvain Reynaud will work on the middleware and software development part. Hélène and Sylvain were absent at the time of the meeting, their work will be discussed next time. In the mean time, it is important to understand what has been already done elsewhere in this field.

Contact with FNAL

Ruth Pordes is now the Executive Director of the OSG project (Congratulation !). The FNAL contact for IGTMD is now Don Petravick.

Dominique will send the minutes to FNAL and will investigate how to start the collaboration with FNAL. One possibility is to make a short presentation at the Wide Area Working Group bi-weekly meeting. The next meeting is on March 10 at 9 am Chicago time (4 pm French time).

Next meeting

April 10, 2006

Action Items:

- 2-27-06 – Pascale: to complete network card selection
- 2-27-06 – Pascale and Jérôme: to define network server configuration
- 2-27-06 – All : send Post-Doc description to relevant places
- 2-27-06 – Jérôme: to explore possibilities to extend a dedicated connection from New-York to FNAL – This should be done in coordination with FNAL network people
- 2-27-06 – Philippe: to explore possibilities to directly connect CC-IN2P3 to New-York and to propose a planning
- 2-27-06 – Pascale: to hire an engineer
- 2-27-06 – CC-IN2P3: to check what is the status of interoperability between EGEE and OSG and identify subjects where the work should be concentrated.

Post-Doc offer:

Key words 10 Gigabit, Wizard Gap, Network diagnosis, High performance transport.
Duration : 2 years
Tutor : Pascale Vicat-Blanc Primet(1) - **email :** Pascale.Primet@ens- lyon.fr.
Position : (1) PhD - HDR - Responsable de l'équipe RESO - Directrice de Recherche INRIA
Laboratoire d'accueil: LIP Ecole Normale Supérieure de Lyon
Working location : Laboratoire IN2P3 - CC Lyon

Goals
This study aims at exploring approaches that may enable to automatically identify, diagnose and correct performance boottlenecks in the context of 10Gb/s very long distance transport.
Abstract
Packet switched and optical transmission networks are potential candidates for extending today's IP- based best-effort service environment. GRID applications can be viewed as an

excellent example for emerging usage scenarios future high performance networking infrastructures have to face with(1 , 2 3 , 4).
However, it is not very clear how the scientific applications that really need it will be able to fully exploit these networking capacities. The main goal of this work is to explore approaches that may enable to automatically identify, diagnose and correct performance boottlenecks in the context of 1Gb/s and 10Gb/s very long distance networks.

The wizard gap is the distance between average network performance and what you really should be able to attain. Back in 1999, Matt Mathis of the Pittsburgh Supercomputing Center first described the existence of a "wizard gap" and predicted that it would grow rapidly. By his estimation, the difference in performance for the average high-end host configured by an average user and one tweaked out by a network expert is somewhere around a factor of 1000 ! The majority of the performance issues are generally attributed to the last 100 meters and the end-hosts. NICs, drivers, TCP settings, application and OS configurations, port negotiation... in their default form, can easily pose serious performance degradation. As the 10Gb/s technology is rapidly coming to our labs and to grid environments running high end scientific applications, it is necessary to examine carefully this problem before the factor increases by more 10 000 ! The disk to disk aspects of end to end data transfers will be also examined and new optimisations in this area, based on the work the INRIA RESO team is doing in high performance cluster networks, will be proposed and evaluated.
This two-year of post-doctoral position will be divided in several steps :

- The goal of the first task will be to survey and evaluate the existing propositions for crossing this gap. A first evaluation campaign will be done on the DataGRID explorer and Grid5000 testbeds , then on a transcontinental 10Gb/s links. As the 100 meter are considered to be the main source of problems, a particular emphasis will be put on end host and LAN diagnosis and optimisation. Bulk data transfer services and high speed transport protocol state of the art will complete this work. Existing software tools, research proposals and standardization efforts in the area will be evaluated and analysed. Of particular interest will be the study and evaluation of solutions proposed by the e2epi Internet 2 group.
- Formulation of requirements and analysis of the applicability of ongoing efforts to the 10Gb/s context will be the second task.
- To explore the specific case of 10Gb/s long distance connections for bulk data transfers in the context of the LHC applications, the third task will concentrate on the specification and the architectural design of a dedicated solution for high energy physics applications (bulk data transfers between tier1 and tier 2 of the Monarch model). The main issues to be solved will be to define generic mechanisms, based on advances in high performance networking, that realize the required performance and inter- operability.
- Validation with scenarii executed in a large scale grid emulated environment.

This work will be done in the IN2P3 lab, in collaboration with INRIA RESO team, RENATER, Fermilab and in the context of the ANR IGTMD project.

1 F Berman, G Fox, and T Hey. The Grid : Past, Present and Future. Wiley, 2003
2 Teragrid : build and deploy the world's largest, most comprehensive, distributed infrastructure for open scientific research.
http ://www.teragrid. org.
3 Datatag : Research & technological development for a transatlantic grid. http ://www.datatag.org

4 Global grid forum : promote and support the development, deployment, and implementation of grid technologies and applica-
tions. http ://www.ggf. org/.
I

Abstract of the project

Grids are high performance and large scale distributed computational and storage systems, used by large user communities. Today many operational grids and grid middleware exist like UNICORE , ARC , EGEE/LCG/gLite/NorduGrid , Globus , Condor , SRB. The emergence and the important deployment of different middleware raise the interoperability problem. These software environments propose global services for job management, data management, access control, status information collect and retrieval. The Global Grid Forum develop standard like OGSA to make these services interoperable, but the set up of the global interoperability requires more research works. On another hand, no standard has been defined for bulk and reliable data transfers in grids.
The aim of this project is to design, develop and validate mechanisms that concretely make the interoperability of heterogeneous grids a reality. The project concentrates on the following topics:

1. Bulk data transfers
2. Replication and referring mechanisms
3. Information system and job management interoperability
4. Grid control and monitoring
5. Usage of statistics and accounting data.

A particular emphasis will be put on disk to disk bulk data transfers over very long distance with optimal performance. The key idea is to fully exploit the specificity of LCG applications and their real infrastructures to analyse and experiment new communication and replication models, alternative transport protocols emerging within the international scientific community. The participation to a standardization process for a generic grid transport service for bulk exchanges between heterogeneous grids will be a strong goal of the project.

Despite the fact that the interoperability and the unification of a generic data transport in Grids are very often perceived as a necessity, they are in fact very little studied. The present project would allow France to get a leading position in this computing area that will be absolutely crucial to insure the LHC data exploitation.

The very strong experience of the partners in deployment and exploitation of international research and production computing instruments gives a promising perspective to this project and its ambitious experimental approach.