

Observations on Using Chimera with LCG-1 Sites Supporting ATLAS

Jerry Gieraltowski, Argonne National Laboratory, PPDG
Rob Gardner, The University of Chicago, iVDGL
Jens Vöckler, The University of Chicago, GriPhyN
January 20, 2004

1. Overview

As part of the Grid2003/Grid3 Project, an ATLAS grid-enabled application package GCE-Server, which is based on GriPhyN Virtual Data System (VDS), was installed on 22 Grid3 sites. Client hosts (GCE-Client) were installed outside Grid3 for job submission. More than 5000 jobs (Geant3 based simulation followed by reconstruction) were processed on about 18 large computing sites, with total data I/O of about 1.1 TB. The data processed were Higgs and Top samples, all registered in a Globus-based RLS server at Brookhaven National Laboratory. A parallel activity was to submit jobs to the LCG-1 production testbed using GCE (Chimera). The submit sites were chosen to be Grid3 sites. The LCG-1 sites were identified to be one site at CERN, one site at Torino, and one site at Brookhaven. This report summarizes our observations based on initial testing efforts. This work is still on-going, and several concrete steps have already been taken to incorporate differences between the LCG and typical Grid3 site configurations.

2. LCG Architecture Considerations

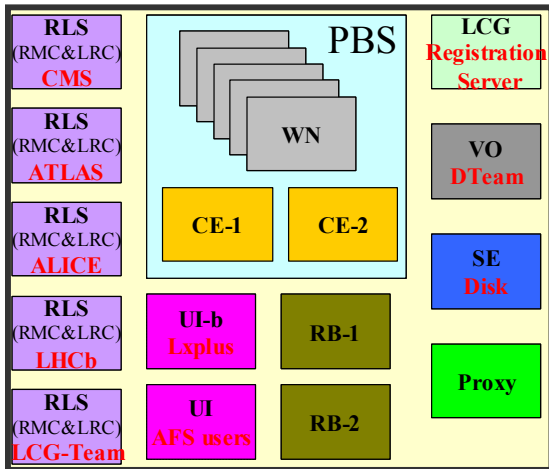
[See [LCG-1 User Guide](#) for more detailed descriptions of architecture components]
As of 11-dec-2003, twenty-five (25) LCG-1 grid gatekeeper servers could be identified as supporting the ATLAS VO. Of these twenty-five, four (4) servers were defined as being part of LCGWEST, nine (9) servers were part of LCGEAST, and twelve (12) servers were part of LCGSOUTH.

The following figure depicts the types of servers that could be found at an LCG-1 site. The servers of interest for this testing effort were:

CE – Compute Element grid gatekeeper

WN – Worker Node

SE – Storage Element



The two sites used in this test effort were “bnllcg1” and “cernlg1”.

The site “bnllcg1” has the following configuration of interest for this testing effort:

CE – atlasgrid04.usatlas.bnl.gov

WNs – two (2)

SE – atlasgrid03.usatlas.bnl.gov

The job schedulers supported on the CE are: “fork” and “lcpbs”

The site “cernlg1” has the following configuration of interest for this testing effort:

CE-1 – adc0015.cern.ch

CE-2 – adc0018.cern.ch

WNs – nineteen (19)

SE – adc0021.cern.ch

The job schedulers supported on CE-1 are: “fork” and “lcpbs”

The job schedulers supported on CE-2 are: “fork” and “pbs”

The following configuration specifics need to be noted:

- An execution mount point for ATLAS (/flatfiles/SE00/atlas) is shared between the WNs and the SE only. This area is not visible to the CEs.
- Gridftp service is available on each WN. Each WN has outbound connectivity within a specific range of ports.
- LCG -1 is installed with VDTALT1.1.8-13 + LCG-specific patches.

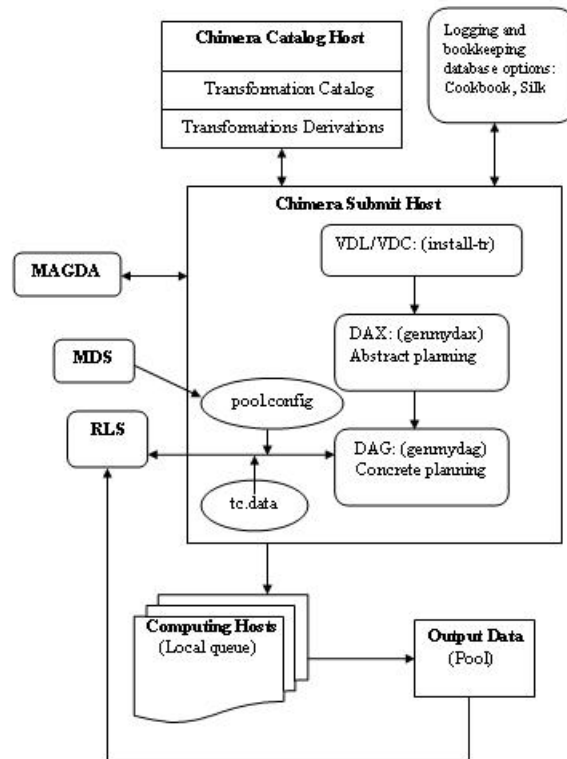


Figure 1 A Grid Component Environment (GCE) use for the Chimera reconstruction data challenge.

3. Initial Testing Results

The following email was sent out on November 19 after some preliminary success was obtained in using Chimera to run an ATLAS simulation job on an LCG-1 site.

“Yesterday I was able to submit a Chimera/Pegasus job to process 4 events using atlsim in the ATLAS 6.5.0 release at the LCG-1 CERN site (adc0018.cern.ch). The job ran to completion, generating the expected “.his”, “.log”, and “.zebra” files. The .log and .zebra files were then successfully transferred to an output pool site at the University of Chicago (grid02.uchicago.edu). The “.his” file failed to transfer. This failure is under investigation. Both the “.log” and “.zebra” files were successfully registered in the RLS at the University of Chicago. A lot has been learned about how LCG-1 sites operate and what type of interactions can occur with non-LCG-1 sites. As expected, several minor script changes were needed to get the software to execute correctly in the lcg1 environment. A number of Chimera/Pegasus enhancements will, undoubtedly, be needed to support a uniform job submission to either an lcg1 site or a non-lcg1 site. The Chimera/Pegasus folks are working directly with Rob and Flavia to achieve this aim. The next step will be to make a few simplifications based on what I have learned already and submit longer running jobs (100 events -> 10 hours), transferring the output files to Brookhaven and registering them in the RLS at Brookhaven.” - Jerry Gieraltowski – ANL

This result, however, proved to be a “lucky” fluke in that the same worker node was chosen by the grid gatekeeper for the stage-out of the output files. More work is

currently in progress by Jens Vöckler (University of Chicago) and Gaurang Mehta (Pegasus team at ISI) to make Chimera/Pegasus perform uniformly in either an LCG or Grid2003 environment.

4. Specific Observations

- Visibility of Working Directories and Executables – Assumptions made by Chimera/Pegasus which were valid in the Grid3 environment proved not to be true in the LCG environment. Specifically, the assumptions that a “working directory” and the actual executable were visible (i.e., shared) to all nodes are not true in an LCG environment. The fact that the grid gatekeeper (CE) can not “see” the storage element (SE) and that the use of pooled accounts results in varying logins on the worker nodes between successive jobs caused a number of problems; some of which have yet to be surmounted.

In-Progress: These items are currently being addressed by the Chimera/Pegasus team.

- Missing LCG Certificate in VDT Release - The CERN LCG site uses a “host certificate” that is not included with any VDT release through VDT-1.1.11. Without the correct CA files associated with this certificate, jobs can not be submitted to the CERN LCG site.

Resolution: VDT support ticket #214 documents this problem. Alain Roy, of the VDT Team, incorporated the necessary certificate files for the missing certificate (fa3af1d7) in the upcoming release – VDT 1.1.12. Private patches of this certificate have been installed on servers at Chicago and ANL to support additional testing with this site.

- Missing Port Assignments for Worker Node Out-Going Connectivity - The distributed list of ports to be assigned and opened on the various LCG nodes does not mention “outbound connectivity” ports on the WNs. The Brookhaven LCG-1 site could not support gsiftp data transfers to the WNs but other LCG-1 sites could. The actual range of ports to be open on the WNs is suggested to be 20000,25000 but the actual port range is left up to the local system admin. Some LCG-1 sites have defined a different range.

Resolution: The Brookhaven system admin (Jason Smith) has arranged to have a specific port range opened on their worker nodes. These connections support outbound connectivity only.

- Correct Environment Setup Needed When Executing on Worker Nodes - Several Globus specific environment variables such as GLOBUS_LOCATION, LD_LIBRARY_PATH, and GLOBUS_TCP_PORT_RANGE are pre-defined for the user prior to execution on a worker node. However, it is not always true that the user can expect the variable PATH to be correctly defined to include Globus binaries.

Resolution: The solution to this problem is left to the user to employ. If your executable is expecting to use Globus commands, you must execute:

```
source $GLOBUS_LOCATION/etc/globus-user.SHELL
```

(where SHELL is either sh or csh)

before executing any other commands in your executable.

- Intentional Host Blockage for GSIFTP Service - The domain name of a worker node is not always published outside of the host site. If this condition is true and a site being connected to from such a worker node is intentionally blocking hosts recognized by a specific domain, the connection will be blocked.

Example: Worker Node atlasgrid24.usatlas.bnl.gov has a numeric DNS designation of 139.66.44.2. The DNS of 139.66.44.2 is recognized as “atlasgrid24.usatlas.bnl.gov” only with the bnl site. If this worker node were to make a request to an external server which is allowing only connections from the domain “.bnl.gov”, the connection would be blocked since the external server can not translate 139.66.44.2 into any recognizable published domain.

Resolution: Servers which expect to interface to worker nodes at LCG sites must not block any domains. The only servers that would fall in this category, from the ATLAS perspective, would be database servers and gridftp servers. One server at Chicago and two servers at ANL have been setup to allow all hosts to connect to them.

- File Permission Problems Due to Pooled VO Account Associations - Login accounts associated with a specific certificate and VO are dynamically assigned from an associated pool of accounts. Thus an ATLAS user may be associated with the account “atlas04” for one job and “atlas06” for the next subsequent job. The user must ensure that file permissions are correctly set for all created files which are intended to remain on the site for subsequent jobs.

Resolution: The user must ensure that created files are group writable. It is suggested that you execute “umask 002” as part of your executable if you are expecting to create any output files which must be available for subsequent jobs.

- Missing PBS Patches from Globus - PBS patches, from Globus, to correctly stage stdout and stderr do not seem to be present in the current LCG-1 versions of the “pbs” and “lcpbs” schedulers. Additional patches to fix known problems with “shared file systems” also do not seem to be included. These fixes have been identified by Jens Vöckler to be distributed with patches to VDT-1.1.12 and are identified by Globus Bug # 950.

In-Progress: Fixes for Globus Bug #950 need to be incorporated into the current LCG release of VDT. *Note 1:* LCG is currently deploying VDTALT1.1.8-13 with additional patches specific to EDG/LCG. It is not clear how dissimilar the two releases, VDT-1.1.12 and VDTALT1.1.8-13, are. *Note 2:* The LCG job scheduler “lcpbs” is not well understood by the authors so it is difficult to determine what has been modified and/or enhanced in this scheduler. Additional documentation from the authors of this scheduler is needed to understand the differences between this scheduler and the standard “pbs” scheduler.

5. Proposed Enhancements to Chimera/Pegasus

- Allow 3rd-party transfers from the submit host to the storage element on the compute host. An attribute will have to be created identifying all hosts which allow third party transfers.
- Modify Pegasus code to automatically ship the kickstart executable to the compute host.
- Modify Pegasus code to remove the dependency on “remote_initialdir” for LCG sites.

6. Current Status of Testing

The Chimera/Pegasus enhancements outlined in the previous section were delivered by Jens Vöckler (GriPhyN) and Karan Vahi (ISI) as a private binary on December 19, 2003. This binary was incorporated into the GCL:GCE-Client package and tested on January 08 and 09. Runs using ATLAS Simulation code based on ATLAS Release 6.5.0 were executed on three (3) different LCG sites; Brookhaven, CERN, and Torino.

All runs were successful with the output data staged to a storage area at The University of Chicago and successfully registered in the Globus RLS at The University of Chicago.. As of 09-jan-2004, additional testing continues with multiple parallel jobs and additional LCG sites.

Site	Processor (MHz)	Number of Events	Elapsed (wall clock time) [hh:mm:ss]	Minutes/Event
Brookhaven	399.33	4	01:11:35	17.9
Torino	2395.95	4	00:09:15	2.3
CERN	796.55	69	05:15:40	4.6

7. Additional Consideration for Interoperability Access

Another possibility to enhance interoperability would be to provide a gatekeeper service on each UI with the only supported job scheduler being “fork”. This would then provide a contact point between the LCG environment at the site and any other Globus-based environment external to the site. Some minor script development would be needed to translate any external job request into an LCG-specific job submission request.

8. Acknowledgements

The authors would like to thank Markus Schulz and Flavia Donno for their untiring support of this testing effort. Special thanks to Markus for his timely explanations of LCG architecture constraints.