GSIMF: A Web Service Based Software and Database Management System for the Next Generation Grids

Nanbor Wang, Balamurali Ananthan

Tech-X Corporation, Boulder, CO – 80301

{nanbor, bala}@txcorp.com

Gerald Gieraltowski, Edward May, Alexandre Vaniachine

Argonne National Labs, Argonne, IL – 60439

{jerryg, may, vaniachine}@anl.gov

Abstract: To process the vast amount of data from high energy physics experiments, physicists rely on Computational and Data Grids; yet, the distribution, installation, and updating of a myriad of different versions of different programs over the Grid environment is complicated, time-consuming, and error-prone.

Our Grid Software Installation Management Framework (GSIMF) is a set of Grid Services that has been developed for managing versioned and interdependent software applications and file-based databases over the Grid infrastructure. This set of Grid services provide mechanism to install software packages on distributed Grid computing elements, thus automating the software and database installation management process on behalf of the users thus enabling users to remotely install programs and tap into the computing power provided by Grids.

Keywords: Computational Grid, Data Grid, Software Management Automation.

1. Introduction

Experimental data from high-energy physics and nuclear physics experiments are commonly stored in mass storage devices for later analysis. The amount of data produced by the experiment devices poses a serious challenge for subsequent offline data analysis like, the ATLAS collaboration [1], which will collect data from the LHC [2] at CERN, which will generate 1-10 petabytes of data a year when it goes into production.

Typically, before final measurements and searches for new physics can be done, each experiment develops a set of specific software programs to perform a series of processing steps on the data collected. Physicists therefore need to install their experiment specific software into the Grid environment before they can start testing the software, making pilot runs, and processing sets of data they wish to analyze. Unlike local computing resources, systems on a Grid are geographically distributed and belong to different organizations, and hence have divergent system hardware and software combinations, as well as configurations, usage and security policies.

Maintaining software installations for systems on a Grid is therefore not a trivial task and is often done manually, which is error-prone and time-consuming. Furthermore, Grids are often shared by multiple groups of researchers.

What is needed is a software installation management system so that an experiment is not hindered from using an otherwise available resource simply because the necessary version of a software package is not present. To solve this, we've modelled our Grid Software Installation Management Framework (GSIMF) as a set of Grid Services for managing versioned and interdependent software applications and file-based databases over the Grid infrastructure on behalf of users.

We have developed key prototype Grid services for querying available software packages and installing software on distributed Grid computing elements. We have demonstrated the prototype services using the ATLAS analysis software package and explored and investigated other software and databases installation management strategies. The following sections report on the implementation details and the functionality of the different Grid services modules and the studies of the applicability of GSIMF for different software installations and the management of various database releases.

2. Architecture

The GSIMF architecture is shown in Figure 1. The description of the core components and the sequence of actions are explained below.

2.1 Roles of each component

2.1.1 Grid Software Package Repository (GSPR):

Software packages are stored as compressed bundles on one or more software package repositories. Since software installation targets on a Grid can belong to different physical organizations with different security policies and firewall configurations, it is important to make the software package repositories accessible by all the potential installation targets. The software repository that we maintain in our Tech-X Corporation Grid is available through GridFTP, so that the software packages that we maintain are easily accessible and installed in other Grid sites that we collaborate with.

2.1.2 Grid Software Catalog Service (GSCS) / Grid Software Package Descriptor Repository (GSPDR):

The Grid software catalog service specializes in keeping track of the metadata of all the software packages made available to a Grid. This software metadata is stored in the GSPDR, catalogues the metadata information describing a software package, such as:

- Name of the software package
- List of directly dependent software packages
- Cache where the package is located
- Mechanisms to retrieve the software package
- Target OS/platform combinations
- Installation history
- Other resources required to installing the software.

The client browses the software catalog and selects the software of interest to install it in the Grid.

2.1.3 GSIMF GUI Client

A detailed explanation of the GSIMF Java GUI client and its mode of operation with screen shots are provided in section 3.

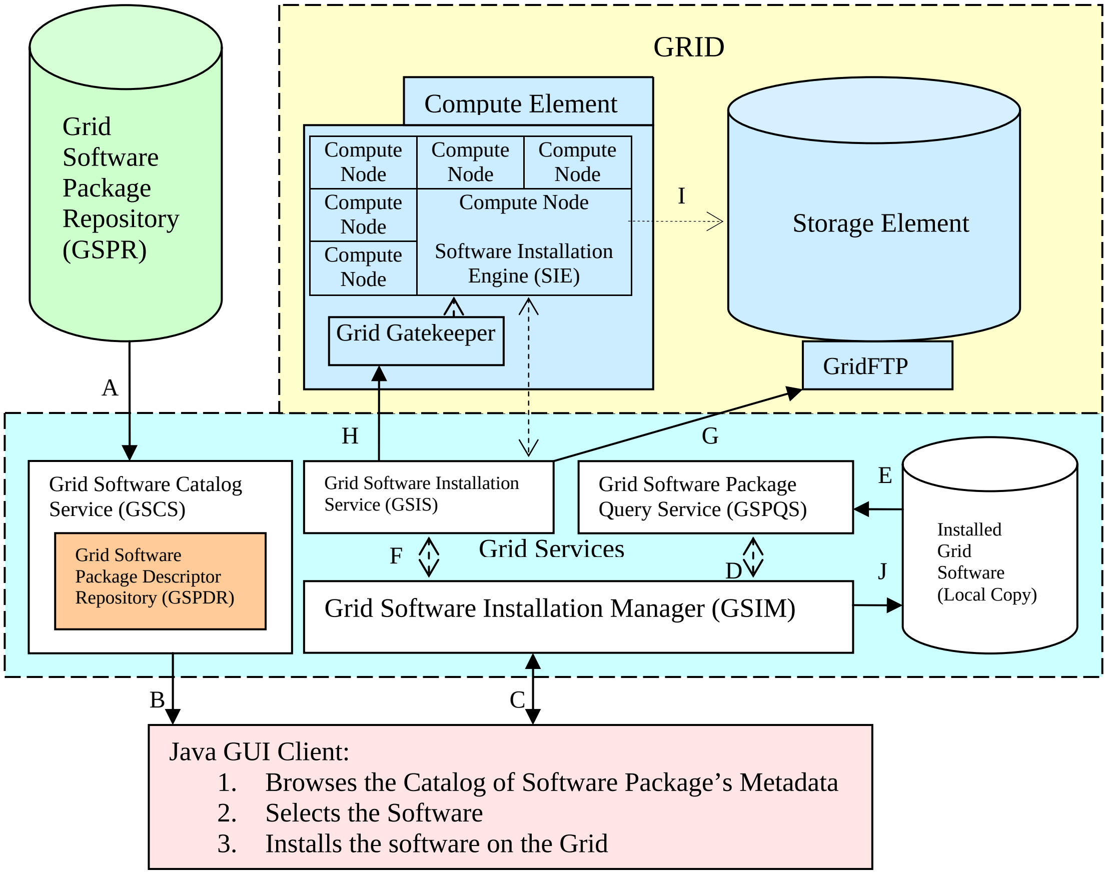

GRID
Grid Software Package Repository (GSPR)
Compute Element
Compute Node
Compute Node
Compute Node
Compute Node
Compute Node
Compute Node
Software Installation Engine (SIE)
Grid Gatekeeper
Storage Element
GridFTP
I
A
H
G
E
F
D
J
B
C
Grid Software Catalog Service (GSCS)
Grid Software Package Descriptor Repository (GSPDR)
Grid Software Installation Service (GSIS)
Grid Software Package Query Service (GSPQS)
Installed Grid Software (Local Copy)
Grid Services
Grid Software Installation Manager (GSIM)
Java GUI Client:
1. Browses the Catalog of Software Package's Metadata
2. Selects the Software
3. Installs the software on the Grid

Figure 1. GSIMF Architecture

2.1.3 Grid Software Installation Manager (GSIM):

GSIM is the central software installation controller that provides services to coordinate overall software installation/removal requests on the Grid. GSIM interacts with the GSIMF client and Grid Software Package Query Service thereby extracting and synthesizing documents that would be used by the Grid Software Installation Service to manage software installations on the Grid. GSIM provides a set of well defined simple interfaces as listed in figure 2:

```
Public interface GSIM{
   int install (PackageInfo softwarePackage);
   String getInstallationResult ();
   boolean uninstall (PackageInfo softwarePackage);
   GetInstalledPackagesResponse getInstalledPackages();
   GetSystemPackagesResponse getSystemPackages();
}
```

Figure 2. GSIM Interface

2.1.4 Compute Element and Storage Element:

The compute element may comprise of many individual machines that form the Grid or a cluster of machines that takes up the responsibility of performing the computational services in the Grid. The storage element is where the software is installed on the Grid with the compute element having accesses privileges to use the installed software. The installed software is then utilized by the Grid jobs that get executed on the compute element.

2.1.5 Grid Software Package Query Service (GSPQS):

This service provides a cataloguing service for all the software packages installed on a storage element on the Grid. This is achieved by maintaining a copy of the installed software with all the dependant software's information on the Grid locally where the GSPQ service runs. Job scheduling and match making services can query GSPQS to decide whether a job can be scheduled on this Grid. When installing a new software package, this service can provide a list of software packages that have already been installed on the CE. Likewise, when removing a software package, this service can provide a list of dependent software packages that are no longer needed. The GSPQS interface is listed in figure 3.

```
Public interface GSPQS {
   Boolean hasPackage(String packageName);
   Boolean hasPackageVersion(String packageName, double versionNo);
   String platform();
   String[] allPackages();
}
```

Figure 3. GSPQS Interface

2.1.6 Grid Software Installation Service (GSIS):

This service accepts software installation or removal requests from GSIM and performs the actions on the end user's behalf. It uses the software metadata descriptor maintained by the Software Package Descriptor Repository for information related to installing a software package on the storage element. GSIS interface is listed in figure 4.

```
Public interface GSIS {
   int install (PackageInfo softwareInfo);
   int uninstall (PackageInfo softwareInfo);
   String installationResult(PackageInfo softwareInfo);
}
```

Figure 4. GSIS Interface

By processing all the software installation services on a Grid through a unique service, we are able to resolve conflicts and avoid any repetitions among multiple installation and removal requests. Internally, the GSIS uses Pacman [3] an existing software installation tool to perform the actual installation, including downloading the software. The external Grid service and internal installation related command line tools are modeled as wrappers [4] around Pacman thereby delegating the actual operations to them.

2.2 Sequence of Actions

The following sequence of actions labelled from A to I is performed by GSIMF system during different course of time. Refer to figure 1 to correlate the actions listed below with the various modules of the system.

A. Grid Software Catalog Service collects information about the software packages stored in Software Package Repository and builds metadata for every individual package and stores in Software Package Descriptor Repository.

B and **C**. The client browses through the catalog of software metadata, selects the software of interest and sends the command to install the software to the Grid Software Installation Manager.

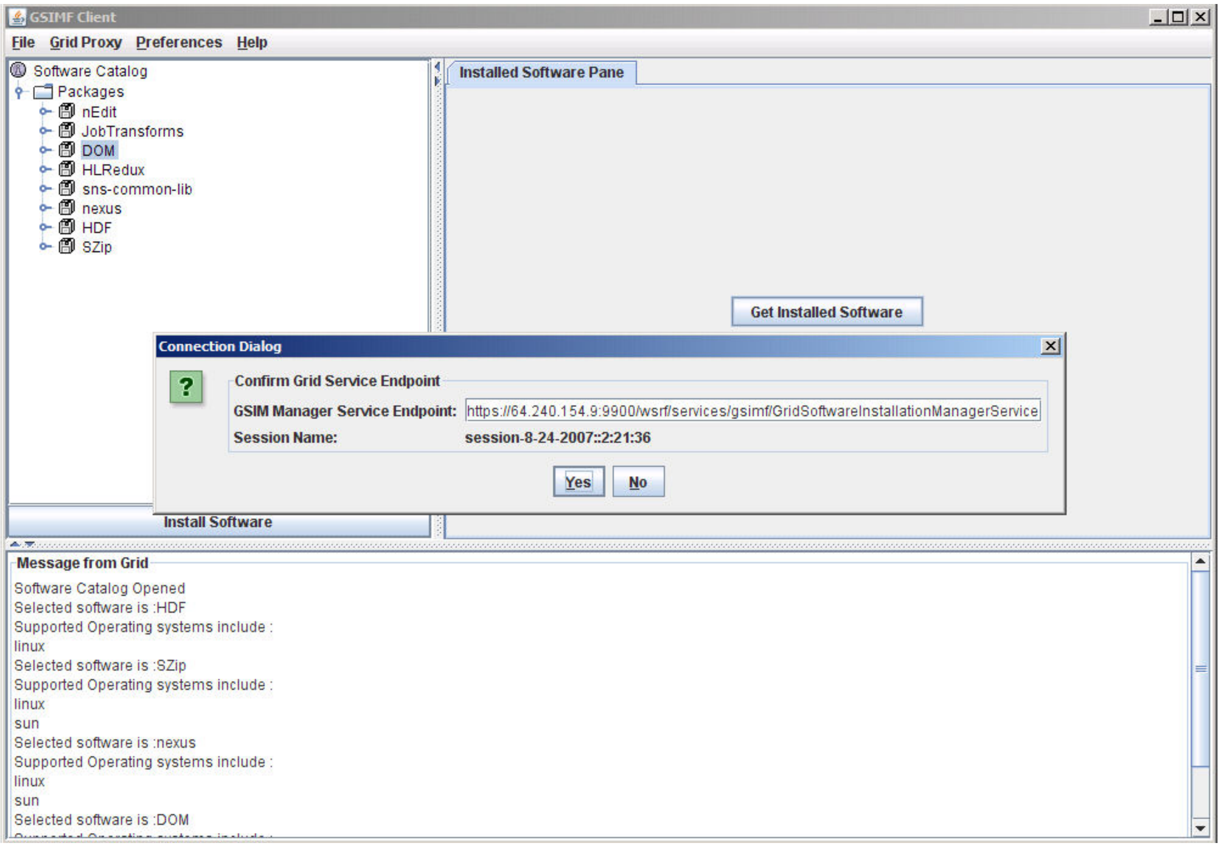
D and **F**. GSIM contacts the Grid Software Package Query Service and after checking that the software is not already installed on the Grid, GSIM delegates the installation processes to the Grid Software Installation Service. If the software is already present, GSIM taken no action and informs the user about the presence of the software. Optionally the user could instruct the GSIM to install a different version of the same software either with or without removing a previous version of the same software.

G and **H**. The GSIS starts a Software Installation Engine on one of the machines in the Grid through the Gatekeeper. The files needed to start the Installation Engine are pre-staged through the GridFTP server. Both the Gatekeeper and GridFTP servers are provided by the Globus Toolkit [5].

I. Software Installation Engine installs the software on the storage element on a location that is accessible by all the nodes of the compute element after taking care of the dependents of the installed software package.

3. GSIMF Java GUI Client

On the Grid, in addition to the software release files, physicists require access to the database-resident data in files. An example of this is the ATLAS Database Release files that are decoupled from the software releases. This task is similar to the installation of the software files performed by GSIMF.



GSIM Manager Service Endpoint: https://64.240.154.9:9900/wsrf/services/gsimf/GridSoftwareInstallationManagerService

Session Name: session-8-24-2007::2:21:36

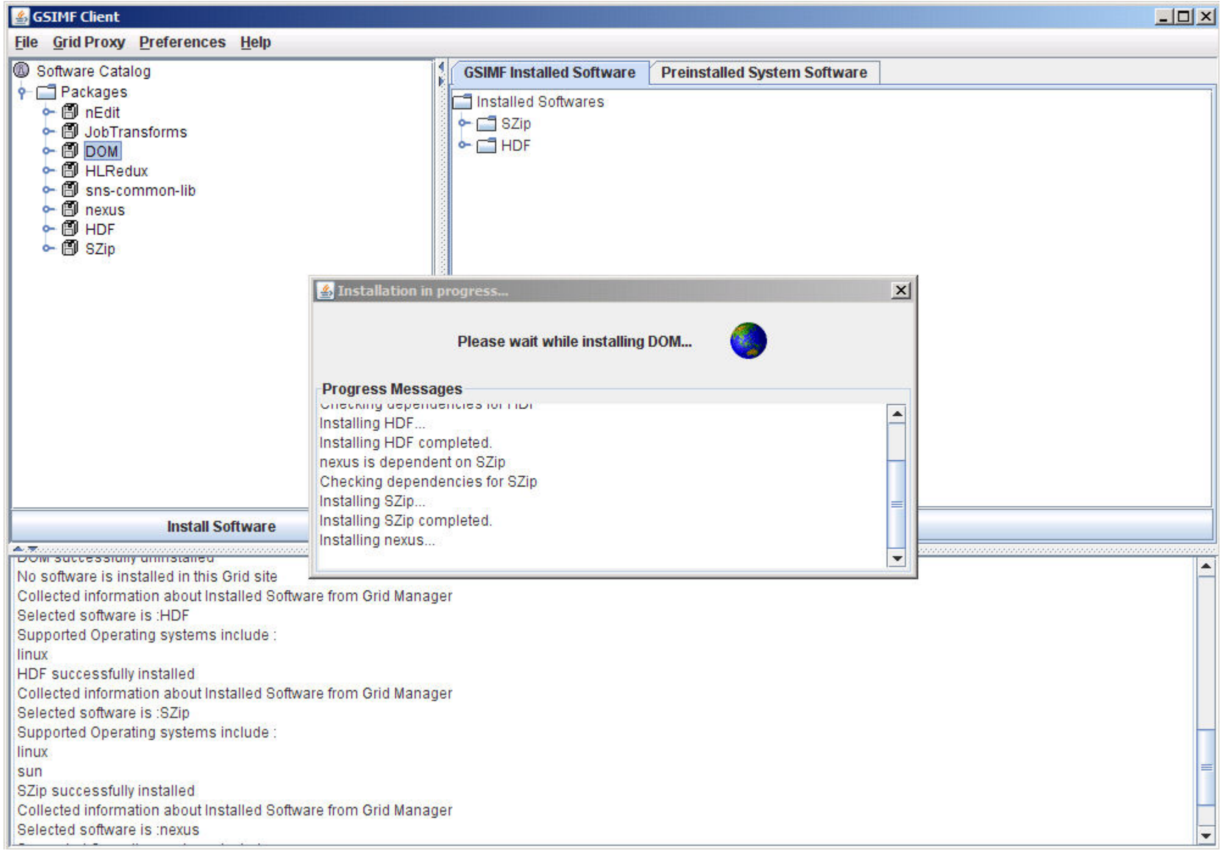
```
Software Catalog Opened
Selected software is :HDF
Supported Operating systems include :
linux
Selected software is :SZip
Supported Operating systems include :
linux
sun
Selected software is :nexus
Supported Operating systems include :
linux
sun
Selected software is :DOM
```

Figure 5. The software catalog is opened on the left panel of the GUI client with which the list of software could be browsed. The user opts to install the DOM software through the GSIM service.

We've developed an interactive Java GUI client that the end user can use intuitively to browse the software catalog and perform software installation and uninstallation on the Grid. Figures 5 and 6 show a couple of screen shots that demonstrates the capabilities of our GSIMF GUI client.

Our GSIMF system is secure as it makes use of the security features available on the Grid Toolkit provided by the Grid Security Infrastructure (GSI) [6]. Provision is provided in the GSIMF client to create Grid proxy certificate that is X.509 [7] certificate complaint. Proxy certificates confirm that the user is authorized by a trusted certificate authority to access Grid resources and the user may delegate their Grid privileges temporarily to another entity to perform tasks on behalf of the user.

All the software management related tasks are performed in the context of a named session which is created by the user initially when he contacts either the catalog service or the GSIM service. When the user is connected with the GSIM service, the list of already installed system software on the Grid is displayed on the 'Preinstalled System Software' tab-pane on the right side panel.



```
Checking dependencies for HDF
Installing HDF...
Installing HDF completed.
nexus is dependent on SZip
Checking dependencies for SZip
Installing SZip...
Installing SZip completed.
Installing nexus...
```

```
DOM successfully uninstalled
No software is installed in this Grid site
Collected information about Installed Software from Grid Manager
Selected software is :HDF
Supported Operating systems include :
linux
HDF successfully installed
Collected information about Installed Software from Grid Manager
Selected software is :SZip
Supported Operating systems include :
linux
sun
SZip successfully installed
Collected information about Installed Software from Grid Manager
Selected software is :nexus
```

Figure 6. GSIMF installing DOM software and its dependants. The list of already installed dependant software is updated on the left side panel with messages from the Grid during the installation is updated on the bottom panel. A popup 'wait' window keeps the user informed with what software is installed at any moment.

Uninstallation of a software too is straight forward in that the user selects the software to be removed and clicks the 'Uninstall' button that sends command to the GSIM service to uninstall the selected software and all its dependants.

4. Conclusion

The GSIMF will contribute to a robust environment for various data intensive and collaborative applications, such as nuclear physics experiments, space science observations, and climate modeling.

The challenge of software installation management in the Grid environment is not unique to the high-energy physics domain. GSIMF can be modified without much effort to bring immediate and similar benefit to other application domains, such as space weather and earth science simulation, and early adopters of OSG such as the search for Gravitational Wave experiment (LIGO) [8], the Sloan Digital Sky Survey (SDSS) [9], and more recently the GeNome Analysis by the GADU group (GNARE) [10]. All these projects require large-scale software systems needing and software installation management tools such as GSIMF which work well and scale well in a GRID environment.

By lowering the cost of deploying and maintaining software on the Grid environment through GSIMF, we can push the envelope of Grid computing technologies further to reach new extents.

5. Acknowledgement

This work is supported in part by the U.S. DOE, division of High Energy Physics, under contract DE-AC-02-06CH11357 and under the U.S. DOE SBIR grant DE-FG02-05ER84384.

6. References

[1] Atlas Experiment. [http://atlasexperiment.org]

[2] The Large Hadron Collider. [http://lhc.web.cern.ch/lhc/]

[3] Pacman Headquarters. [http://physics.bu.edu/pacman]

[4] Gamma. E, Helm. R, Johnson. R, and Vlissides. J, *Design Patterns: Elements of Reusable Object-Oriented Software*. Reading, MA: Addison-Wesley, 1995.

[5] Globus Toolkit. [http://www.globus.org/**]**

[6] Grid Security Infrastructure [http://www.globus.org/security/overview.html**]**

[7] X.509 certificates [http://www.ietf.org/html.charters/pkix-charter.html]

[8] LIGO. [http://www.ligo.caltech.edu/LIGO_web/about/factsheet.html]

[9] SDSS. [http://www.sdss.org/**]**

[10] GNARE. [http://compbio.mcs.anl.gov/gnare/doc.cgi]