GridAnt – Client Side Grid Workflow Management with Ant

Gregor von Laszewski,[1,*] *Beulah Alunkal,*[1] *Kaizar Amin,*[1] *Shawon Hampton,*[2] *Sandeep Nijsure*[1]

Draft Whitepaper

`http://www.globus.org/cog/grant.pdf`

[*] *Corresponding Author: gregor@mcs.anl.gov*
[1] *Argonne National Laboratory, 9700 S. Cass Ave, Argonne IL, 60440*
[2] *National Center for Supercomputing Applications, 605 E. Sprinfield Ave., Champaign, IL 61820*

July 31, 2002, Revised October 24, 2002, January 8, 2003, February 1, 2003

Contents

Abstract

GridAnt provides a simple but extensible framework for rapidly prototyping Grid applications based on XML specifications. Basic tasks such as authentication, file transfer, and job execution are predefined. As the name indicates the implementation of GridAnt is based on the commodity tool *ant*. Hence, it provides the ability to use features from ant such as the XML specifications of tasks, the control flow specification through conditional, sequential, and parallel constructs, and the availability of a workflow processing engine. As GridAnt provides an abstract definition of elementary tasks performed in Grids it is possible to use the same application specification on a variety of Grid hosting environments. We demonstrated at SC'2002 that the same specification can be used on the Globus Toolkit versions 2 and 3.

1 Introduction

The popularity of UNIX and DOS was initiated through the availability of sophisticated shells to express command line sequences. Soon it was noticed that complex workflow dependencies as part of the program generation process needed to be expressed. Makefiles have an essential part in automating the complex build process leading to a revolution in software engineering. It allowed users to structure their code and to build sophisticated reusable libraries with ease.

For the Java Platform a similar tool exists under the Jakarta Project with the name *ant* [1]. It provides the Java community with a sophisticated build tool. Dependencies and parallelism can be expressed easily and new tasks can be included with little effort. As ant is used by a large community it has been transitioned to a commodity tool and is the defacto standard for controlling build processes in Java. By now every major Java interface development environment [2, 3, 5] includes the possibility to generate class files through ant. Based on our experience with Grid workflows [**?**, **?**] we propose to reuse the *ant* framework to develop a simple yet powerful client side workflow framework for Grids, which we named *GridAnt*.

With GridAnt we try to answer several questions:

- Is it possible to base an implementation on ant for simple workflows in Grids?
- Is the language implied by the ant framework to express workflows powerful enough for Grid workflows?
- Can we reuse the developed framework to be integrated in a Grid shell?

These and other questions are addressed by us in a prototype implementation that follows loosely the specification within this paper. We find that the availability of our GridAnt framework provides much needed functionality for testing and developing basic Grid applications. Furthermore, we found that the integration into a testing framework make it more easily possible to develop unit tests for production Grids while being able to monitor the existing Grid infrastructure through test scripts that require dependencies and parallel constructs. Such tests can be accessed through junit [4].

2 GridAnt Tasks

For *GridAnt* to be most useful for the large Grid community, we first identified a set of essential Grid tasks that must be supported. This includes tasks to

setup : the Grid environment (Section 2.1),

authenticate : users to the Grid (Section 2.2),

copy : files between Grid resources (Section 2.3),

delete : files on a Grid resource (Section 2.4),

execute : tasks on the Grid (Section 2.6),

query : information about the Grid (Section 2.8), including the status (Section 2.9) of the submitted tasks,

checkpoint : the current state of tasks to be executed (Section 2.11).

In the following sections we will introduce XML specifications that can be used within GridAnt to issue such tasks.

2.1 grid-setup

The *grid-setup* task is used to set and return basic environment variables used within GridAnt. It has a limited number of arguments that have a global impact on optional attributes within every task. However, the value of each variable can be overwritten locally by a task.

The task *grid-setup* without any parameters returns information about the grid environment in use. This includes the version of the JVM and ant, the values of system variables, and the environment variables set.

Optional Parameters

provider : is a String that sets the default hosting to a particular provider of Grid hosting environments. Values are *GT2* for the Globus Toolkit Version 2.2, *OGSA* for the OGSA based services, and ssh for the Operating system based ssh service. The default value is *GT2*.

debug : is a Boolean, that, if true, switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean, that if, true initiates a dryrun without actually executing the tasks. The default value is false.

log : is a String that specifies a filename in which a log of the executed GridAnt tasks are written (see Section 3 for more details).

Examples

Print the Environment.

```
1 <grid-setup />
```

Set the dryrun, debug, and provider variables.

```
1 <grid-setup dryrun="true" debug="true" provider="GT2"/>
```

2.2 grid-authenticate

The task *grid-authenticate* is used to prepare the system for subsequent authentication as part of a single sign on process by the user to the Grid. In case of a Globus hosting environment a security proxy is created that has limited lifetime and is reused in behalf of the user to authenticate to the system. The authentication mechanism is dependent on the hosting environment. All Grid tasks need this authentication preparation step.

Optional Parameters

userkey : is a String that specifies the location of the *userkey*.

usercert : is a String that specifies the location of the usercert.

authority : is a String and specifies the location of the public key of the certificate authority, the keys are generated with.

directory : is a String that specifies the directory in which the *userkey*, *usercert* and the *authority* are located. In case the *userkey* and/or *usercert* are not explicitly defined and the hosting environment is Globus, the default names *userkey.pem* and *usercert.pem* are use respectively. The default filename in which the public key for the signing authority is specified is the first file in alphabetical order in the directory that has a name ending with ".0".

visual : is a Boolean that if set to true initiates a graphical user interface to generate a Grid proxy as part of the authentication process. If set to false an interactive command line is used for authentication. The default value is true. Note, that due to internationalization efforts the spelling "vizual" is also acceptible.

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Examples

Implicit generation of a grid proxy as part of the first step in the Grid authentication process

```
<grid-authenticate visual="true" directory="/home/gregor/.globus" />
```

Implicit generation of a grid proxy as part of the first step in the Grid authentication process

```
<grid-authenticate visual="true"
        authority="/home/gregor/.globus/481234.0"
        usercert="/home/gregor/.globus/usercert.pem"
        userkey="/home/gregor/.globus/userkey.pem" />
```

2.3 grid-copy

The task *grid-copy* copies a file from one Grid resource to another. Third party transfers are supported through the use of the gridftp protocol.

Required Parameters

name : is a String that specifies a name under which the task can be referenced during runtime.

to : is a String that specifies the location where the file is copied to.

from : is a String that specifies the location where file is copied from. In case a directory name is specified all files within this directory are copied. The to location must point to a directory also.

Optional Parameters

server : is a String that specifies the address of the reliable file transfer server which is used to transfer the file.

arguments : is a String that specifies the arguments for the reliable file transfer server which is used during the transfer.

recursive : is a Boolean that specifies if a directory copy is performed recuresively.

timeout : is an Integer that specifies the time in seconds the transfer should not exceed. In case the time is exceeded the transfer is interrupted.

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Examples

Reliable file transfer

```
<grid-copy name="rftcopy"
        server="rft.mcs.anl.gov"
        from="http://www.mcs.anl.gov/~gregor/file.txt"
        to="gridftp://hot.mcs.anl.gov:/home/gregor/to.txt" />
```

Simple copy

```
<grid-copy name="gridtp-copy"
        from="gridftp://www.mcs.anl.gov/~gregor/file.txt"
        to="gridftp://cold.mcs.anl.gov:/home/gregor/to.txt" />
```

Simple copy of a directory tree that copies all files in directry A and their subdirectories to directory B.

```
<grid-copy name="gridtp-copy"
        recursive="true"
        from="gridftp://www.mcs.anl.gov/~gregor/dirA"
        to="gridftp://cold.mcs.anl.gov:/home/gregor/dirB" />
```

2.4 grid-delete

The task *grid-delete* deletes a file from a Grid resource.

Required Parameters

name : is a String that specifies a name under which the task can be referenced during runtime.

file : is a String that specifies the location of the file to be deleted.

Optional Parameters

timeout : is an Integer that specifies the time in seconds the transfer should not exceed. In case the time is exceeded the transfer is interrupted.

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Examples

Delete a file

```
<grid-delete name="delete-file"
        file="gridftp://hot.mcs.anl.gov:/home/gregor/to.txt" />
```

2.5 grid-mkdir

The task *grid-mkdir* creates a new directory on a Grid resource.

Required Parameters

name : is a String that specifies a name under which the task can be referenced during runtime.

directory : is a String that specifies the location of the directory.

server : is a String that specifies the address of the reliable file transfer server which is used to transfer the file.

protocol : is a string that specifies the protocol which is used to connect to the server. Valid protocol values are GRAM and gridftp.

Optional Parameters

timeout : is an Integer that specifies the time in seconds the transfer should not exceed. In case the time is exceeded the transfer is interrupted.

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Examples

Create a directory

```
1 <grid-mkdir name="makedir"
2         protocol="gridftp"
3         server="hot.mcs.anl.gov"
4         directory="/home/gregor/dir" />
```

2.6 grid-execute

The task *grid-execute* executes a command on a Grid resource.

Required Parameters

name : is a String that specifies a name under which the task can be referenced during runtime.

Optional Parameters

directory : is a String that specifies the directory on the resource in wich the command is executed.

environment : is a String that is passed to the command to set up the environment variables.

timeout : is an Integer that specifies the time in seconds the transfer should not exceed. In case the time is exceeded the transfer is interrupted.

output : is a String that specifies the location to which the output should be written.

error : is a String that specifies the URI where the error should be stored.

protocol : is a String that specifies the protocol and the version of the hosting environment.

server : is a String that specifies the server in hostname:port format or the factory contact in case it is a GT3 hosting environment.

timeout : is an Integer that specifies the time in seconds the transfer should not exceed. In case the time is exceeded the transfer is interrupted.

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Example

```
<grid-execute name="ls-hot"
        environment="PATH=/bin"
        directory="/home/gregor/project"
        protocol="GRAM"
        server="hot.mcs.anl.gov:2135"
        executable="/bin/ls"
        output="dir.txt"
        timeout="60s"
/>
```

2.7 grid-cancel

The task *grid-cancel* cancels the job with a given name.

Required Parameters

name : the name of the task to be canceled

Optional Parameters

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Example

```
1 <grid-cancel name="ls-hot">
```

2.8 grid-query

A grid query is used to obtain information about the Grid environment. In previous sections we have already seen some elementary query functions such as *grid-ps*, *grid-environment*, and *grid-status*. In contrast to these specialized functions *grid-query* provides a general interface to queries issued against Grid Information services.

name : is a String that specifies a name under which the task can be referenced during runtime.

server : is a String that specifies the server in hostname:port format.

parameter : is a String and specifies the query to be issued against the specified server.

protocol : is a String that specifies the protocol the server is speaking. Protocols can be LDAP and SQL.

Optional Parameters

timeout : is an Integer that specifies the time in seconds the transfer should not exceed. In case the time is exceeded the transfer is interrupted.

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Example

An information query for an MDS service

```
1 <grid-query name="query"
2         server="mds.globus.org:2135"
3         parameter="(objectclass=*)"/>
```

2.9 grid-status

The task *grid-status* returns the status of a named task. In case no parameter is used it returns the status of all tasks.

Optional Parameters

timeout : is an Integer that specifies the time in seconds the transfer should not exceed. In case the time is exceeded the transfer is interrupted.

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Required Parameters

name : the name of the task to request its status.

Example

Request the status of a named task.

```
1 <grid-status name="name"/>
```

Request the status of all tasks.

```
1 <grid-status />
```

2.10 grid-ps

The task *grid-ps* lists the processes running on the specified grid hosting environment.

Required Parameters

name : is a String that specifies a name under which the task can be referenced during runtime.

server : the specification of the server

Optional Parameters

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

```
1 <grid-ps name="ps"
2          server=pitcairn.mcs.anl.gov:2135"/>
```

2.11 grid-checkpoint

The task *grid-checkpoint* writes a checkpoint to the specified file. Checkpointing does not checkpoint remote tasks but only client side GridAnt tasks. This allows the client to be restarted at another time.

Required Parameters

file : is a String and specifies the file name in which to write the checkpoint.

Optional Parameters

interval : is an Integer that specifies a time interval in seconds between consecutive checkpoints.

overwrite : is a Boolean that specifies if the previous checkpoint should be overwritten.

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

Example

Checkpointing GridAnt:

```
1 <grid-checkpoint file="file://home/gregor/checkpoint.txt"/>
```

2.12 grid-restart

Restarts from the checkpoint specified.

Required Parameters

file : is a String and specifies the file name from which to read the checkpoint.

Optional Parameters

debug : is a Boolean that if true switches on a debugging mode. In the debug mode verbose information is printed during the execution of GridAnt scripts. The default value is false.

dryrun : is a Boolean that if true initiates a dryrun without actually executing the tasks. The default value is false.

```
1 <grid-restart file="file://home/gregor/checkpoint.txt"/>
```

3 Log File

Each task produces two log entries into the log file specified with the *grid-setup* task. The first log entry is created at the task startup, while the second entry is created whne the task is completed. An entry in the logfile are specified in XML.

```
1 <timestamp time="10:15:15 01/02/2002" action="start" >
2 ... task ...
3 </timestamp>
4
5 ... elapsed time ...
6
7 <timestamp time="10:23:25 01/02/2002" action="end" >
```

```
... task ...
</timestamp>

```

4 Control Flow

Ant contains a control construct for expressing parallel and sequential tasks. Taks encapsulated in <sequential> </sequential> are executed in sequential order. Tasks encapsulated in <parallel> </parallel> are executed in parallel.

4.1 Parallel execution

The ant parallel construct allows easy parallel execution of tasks.

```
<parallel>
    <grid-copy from="gridftp://sunny.mcs.anl.gov:/home/gregor/a.txt"
            to="gridftp://cold.mcs.anl.gov:/home/gregor/b.txt">
   <grid-copy from="gridftp://white.mcs.anl.gov:/home/gregor/c.txt"
            to="gridftp://black.mcs.anl.gov:/home/gregor/d.txt">
</parallel>
```

4.2 Sequential execution

The ant sequential tag allows easy sequential execution of tasks.

```
<sequential>
      <grid-execute name="ls pitcairn"
            environment="PATH=/home/gregor/bin"
            directory="/home/gregor/project"
            protocol="GRAM"
            server="white.mcs.anl.gov:2135"
            executable="preprocess"
            output="dir.txt"
      />
      <grid-copy from="gridftp://white.mcs.anl.gov:/home/gregor/c.txt"
            to="gridftp://black.mcs.anl.gov:/home/gregor/d.txt">
      <grid-execute name="ls pitcairn"
            environment="PATH=/home/gregor/bin"
            directory="/home/gregor/project"
            protocol="GRAM"
            server="black.mcs.anl.gov:2135"
            executable="process"
            output="dir.txt"
      />
   <sequential>
```

4.3 Dependencies through targets

Ant provides a nice way of defining dependencies between tasks. The tasks are resolved based on the dependency graph.

```
 <target name="getFileA"
    description="retrieves the File a.txt" >
```

```
    <grid-copy from="gridftp://sunny.mcs.anl.gov:/home/gregor/a.txt"
               to="gridftp://cold.mcs.anl.gov:/home/gregor/a.txt">
</target>

<target name="getFileB"
   description="retrieves the file b.txt" >
   <grid-copy from="gridftp://sunny.mcs.anl.gov:/home/gregor/b.txt"
               to="gridftp://cold.mcs.anl.gov:/home/gregor/b.txt">
</target>

<target name="calculate" depends=''getFileA,getFileB''
   description="calculates when file a.txt and b.txt have been retrieved" >
      <grid-execute name="ls pitcairn"
            environment="PATH=/home/gregor/bin"
            directory="/home/gregor"
            protocol="GRAM"
            server="cold.mcs.anl.gov:2135"
            executable="process"
            output="result.txt"
      />
</target>
```

Acknowledgement

We like to thank Jarek Gawor, Thomas Sandholm, and Ravi Maddhuri for answering our questions we had during the implementation phase of our first prototype released before SC'2002.

This work is supported by the Mathematical, Information, and Computational Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy under Contract W-31-109-Eng-38. The work is conducted as part of the SciDAC CoG Kit project. The Java CoG Kit is developed at Argonne National Laboratory as part of the Globus Toolkit. GlobusToolkit and Globus Project are trademarks held by the University of Chicago.

References

[1] Ant – a Java-based Build Tool. `http://jakarta.apache.org/ant/`.

[2] Eclipse: An Open and Extensible IDE. `http://www.eclipse.org`, 2003.

[3] JBuilder. `http://www.borland.com/jbuilder/`, 2003.

[4] JUnit, Testing Resources for Extreme Programming. `http://www.junit.org/`, 2003.

[5] Sun ONE Studio. `http://wwws.sun.com/software/sundev/`, 2003.

A Implementation

Whenever a task is completed change the value to done.

task	ogsa	cog	notes
grid-environment		Beulah	
grid-autenticate	Kaizar	Beulah	same for ogsa and cog
grid-copy	Kaizar	Beulah	
grid-execute	Kaizar	Beulah	
grid-cancel	Kaizar	Beulah	
grid-delete	Kaizar	Beulah	
grid-status	Kaizar	Beulah	
grid-ps	Kaizar	Beulah	
grid-checkpoint			
grid-restart			

B Document Source

The document source is maintained in LaTeX and is accessible through the Java CoG CVS archive. The tags are "design" and "papers/bib." The latter contains biblioigraphical references.