Nordic Testbed for Wide Area
Computing and Data Handling

NORDUGRID-TECH-1
19/02/2002

Architecture Proposal

M.Ellert, A.Konstantinov, B.Kónya, O.Smirnova, A.Wäänänen

Introduction

The document describes the minimal architecture, necessary to enable a basic Grid functionality, and suggests ways to implement it.
Globus Toolkit™ is used as the basis; all modifications should preserve backward compatibility as much as possible

Terms

BR	Broker
CA	Authentication service (Certificate Authority)
CN	Computing Node (one machine in a cluster)
IS	Information service
MDS	Globus MDS (GIIS and GRIS)
RC	Replica Catalog
RM	Replication Manager
SE	Storage Element
VO	Authorization service
UI	User interface
WN	Worker Node = CN
SD	Session Directory
GM	Grid Manager
GWD	Grid Working Directory

Architecture

In what follows, a description of components is given and their functionality outlined

Cluster

- At present, a PBS cluster
- WN's are not required to access external network
- Application software is installed on the front-end and exported to the nodes; application software installation on WN's is not required
- Recommended shared Grid Working Directory (e.g. /scratch, or /job)
- Services (e.g. data movement, IS) are run only from the front-end

Storage Element

- A separate (stand-alone) machine for flat file storage and/or database server
- Can be "local" to a cluster (i.e. exported disks)
- Runs GridFTP server
- Produces entries to the MDS (if a separate machine, runs its own GRIS)
- Has Grid certificate-based authorization and quotas
- Mirroring and replication is done by services running on the SEs

Front-end

- Authorization and authentication is done by the gatekeeper
- Provides pre- and post-job file/data transfers via GridFTP (runs the server) and/or

symbolic linking in case of local SE

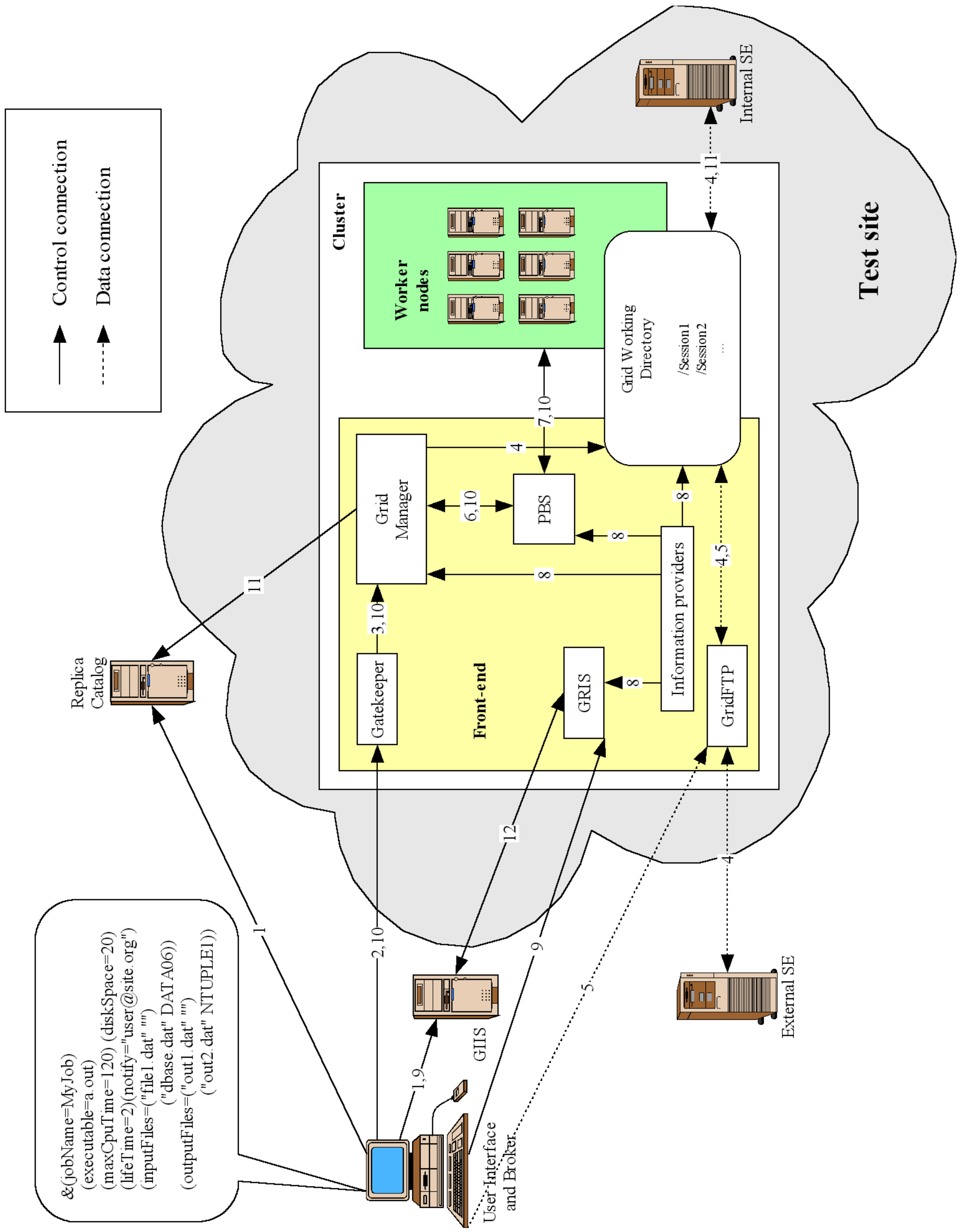
- Provides job submission services, using standard Globus interface
- The Grid Manager (GM) is introduced, which:
  - creates the Session Directory (SD)
  - creates the job ID (to appear in the MDS)
  - pre-stages input files from SE's
  - initializes PBS job submission
  - moves requested output files to SE's and registers them in RC
  - manages job cancellation
  - maintains the set of job status files
  - sends an (optional) e-mail notification to a job owner, both at the start and end of a job (including staging in/out)
- Has a Grid Working Directory (GWD) for stage in/out, shared with the rest of the cluster
- Maintains SD's:
  - a new SD is created in GWD per job
  - only the actual owner of the job can access the SD contents (using the job handle)
  - the SD should be locked for deletion during the course of the job; files should be removed/overwritten only by the GM
  - SD is erased after users' retrieval of its contents; otherwise it is erased after its the pre-defined lifetime expires

Information System

- Based only on Globus MDS
- Hierarchic
- Providers run on clusters and on SEs
- Contains cluster information, both static and dynamic:
  - total/available disk space
  - cluster and jobs status
  - authorization information
- Contains SE information

User interface

- Integrated with the broker (aka decision engine)
- Does not run any server (i.e., client-only)



```
&(jobName=MyJob)
 (executable=a.out)
 (maxCpuTime=120) (diskSpace=20)
 (lifeTime=2)(notify="user@site.org")
 (inputFiles=("file1.dat" "")
        ("dbase.dat" DATA06))
 (outputFiles=("out1.dat" "")
        ("out2.dat" NTUPLE1))
```

Figure 1. NorduGrid task flow (see explanations in Table 1)

1	User Interface queries GIIS and the Replica Catalog, and selects the site; session directory name and the job contact string are created by the Broker
2	Job request: xRSL file is submitted to the Gatekeeper, along with the session directory name and the [resolved by the Broker] physical file names
3	Gatekeeper parses the job request to the Grid Manager
4	Grid Manager: - creates the Session Directory - downloads/links files from Storage Elements to the Session Directory
5	User Interface uploads input files and executables via GridFTP
6	After all files are pre-staged, the Grid Manager submits the job to PBS
7	PBS schedules the job on the Working Nodes
8	Information Providers collect job and queue information, disk usage information and other static parameters (pull); write it to MDS
9	User Interface monitors the job status by querying the MDS
10	User Interface may cancel jobs (Session Directory should be removed)
11	Grid Manager moves requested output results to Storage Elements (initially, Internal only) and registers in the Replica Catalog
12	GIIS queries GRIS

Table 1: NorduGrid task flow (numbering corresponds to Figure 1)

1. Job submission

- Makes use of extended RSL syntax for the job options file
- Performs matching (using the MDS) of job options to a resource, e.g.:
    - required CPU time
    - required system
    - required disk space
    - required runtime environment
    - required memory
    - required data from SE's
- Creates SD name (to be parsed along with the job options)
- Resolves logical file names for input files
- Triggers the GM: parses the job options along with resolved file URL's and the SD name to the matching cluster
- Transfers (GridFTP push) input files and binaries to the matching cluster's SD: either uploads local files, or initializes 3d-party transfer of remote files (or both)

Job monitoring

- Query job status (stored in the MDS) on the cluster via the contact string (contains job handle)
- Capture snapshot of any user-specified file in the SD

Job output

- User initializes download of output files from the stage area

Other features

- Proxy initialization
- List of matching resources (w/o actual job submission, dryrun)
- Job cancellation
- Possibility to submit jobs to explicitly specified resources

Data movement

- All the data are moved before the job submission to the clusters stage area
  - pushed from the user interface machine
  - downloaded from an external SE
  - linked from a local SE
  - job can not request new data sets during execution
  - in the future, a Grid file system can remove such a restriction
- Output is downloaded by the user interface from the SD of the GWD of a cluster, or is written to a SE

Proposed architecture and the task flow is shown schematically in Figure 1 and Table 1.