NorduGrid Testbed: Architecture overview & the Toolkit

Nordic Testbed for Wide Area Computing and Data Handling

NorduGrid Project

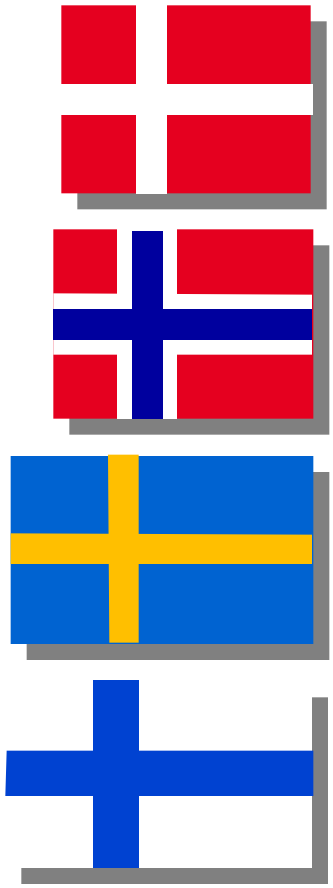
www.nordugrid.org

- Create a Grid infrastructure in Nordic countries
- Operate a production quality Testbed
- Expose the infrastructure to end-users of different scientific communities
- Survey current Grid technologies
- Pursue basic research on Grid Computing
- Develop Middleware Solutions

"preprint" broschure:www.nordugrid.org/documents/booklet.pdf

Nordic Testbed for Wide Area Computing and Data Handling

Participants



Copenhagen University: Niels Bohr Institute, Research Center COM, DIKU

Oslo University, Bergen University

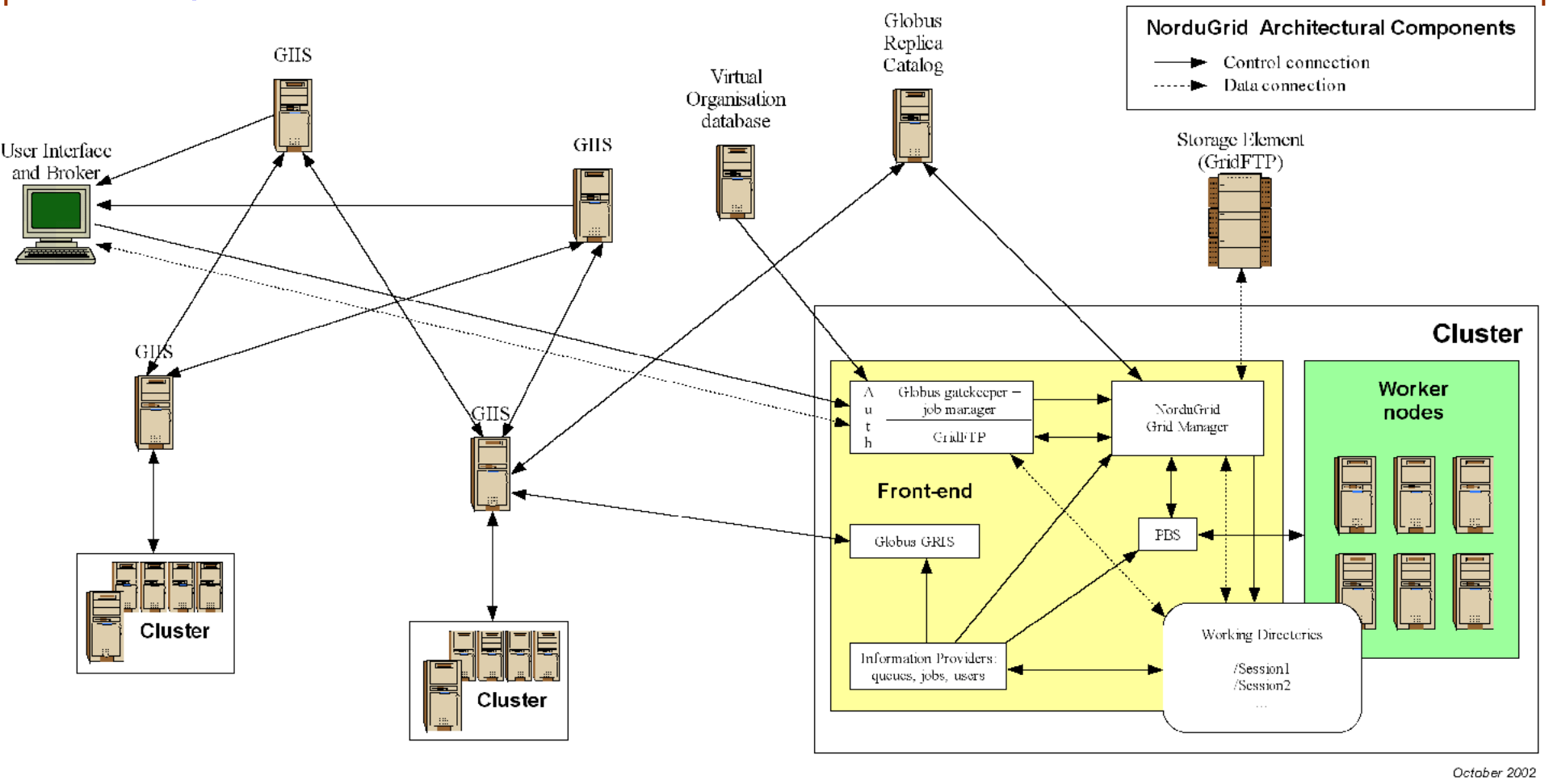
Lund University, Uppsala University, Stockholm University, KTH

Helsinki Institute of Physics

resources:

www.nordugrid.org, and click on the Loadmonitor

architecture



An overview of an architecture proposal for a high energy physics Grid, Lecture Notes in Computer Science 2367, 76 (2002), http://arxiv.org/abs/cs.DC/0205021

Nordic Testbed for Wide Area Computing and Data Handling

NorduGrid Toolkit:

- it is:
  - a functional middleware solution developed by the NorduGrid project
  - implements the fundamental Grid services
  - extends the Globus Toolkit
  - replaces/obsolates some of the Globus core services
- it is not:
  - just a webinterface, a monitoring tool
  - an oversimplified Grid toolkit
  - a complete solution

Nordic Testbed for Wide Area Computing and Data Handling

the components

- Grid Manager (clever stage in/stage out, job management on the cluster)
- GridFtp server (data transfer)
- UserInterface (command line ui + built in broker)
- Extended RSL (job & resource request specification)
- Information Model/System (LDAP-based, job monitoring!)
- Load Monitor (very nice ldap/php based monitoring tool)
- user management (certificate-based VO management)
- very much needed:
  - a reliable data management system, distributed replica management
  - better AAA layer, Grid user management, “Grid access control”
  - GridPortal

Grid Manager

- Provide job control and data handling functionalities
- the middleware layer which sits/runs on top of the LRMS
- job control: submit/cancel jobs by interfacing to the LRMS
- data handling:
  - “stage in” input data and executables either from the UI, SEs, can resolve logical names by contacting an RC
  - “stage out” output data.
  - creates and manages the job's session directory
  - cache management (stores input files in a cache)
  - keep results on cluster untill user downloads.
  - uploads files to the SE, registers them to the Replica Catalog.
  - file transfer is done via the GridFTP server

Nordic Testbed for Wide Area Computing and Data Handling

Grid Manager cont.

- further features:
  - E-mail notification of job status changes.
  - Support for software runtime environment configuration, GM dynamicaly sets the requested Unix environment for the application
- the GM is implemented as a single daemon which uses special GridFTP plugins:
  - certificate oriented local file system access plugin
  - job submission/access plugin
- Limitation:
  - Data is handled only at the beginning and end of the job. User must provide information about input and output data.

Nordic Testbed for Wide Area Computing and Data Handling

UserInterface

- command line tools:

ngsub	- for job submission
ngstat	- to obtain the status of jobs and clusters
ngcat	- to display the stdout or stderr of a running job
ngget	- to retrieve the result from a finished job
ngkill	- to kill a running job
ngclean	- to delete a job from a remote cluster
ngsync	- create a local synchronised copy of the local distributed job information
ngmove	- file transfer

- built-in brokering upon

  user request, “free” resources, required file transfers

Nordic Testbed for Wide Area Computing and Data Handling

UserInterface cont.

- The UI processes user-level xRSL request and transforms to a form suitable for GM
- Performs brokering (built-in Broker)
  - analyzes information about the different clusters obtained from the MDS
  - analyzes information about required file transfer obtained from the Replica Catalogue
  - from all suitable queues one is chosen randomly, with a weight proportional to the amount of free computing resources
- Passes modified job request to GM through GridFTP interface and uploads input files.
- Can be used as an MDS interface for job & cluster status

a brokering session

```
[konyab]$ ./ngsub -d 1 -f ~/gm_test/ui_sleep.rsl
User subject name: /O=Grid/O=NorduGrid/OU=quark.lu.se/CN=Balazs Konya
Remaining proxy lifetime: 5 hours, 1 minute
Initializing LDAP connection to grid.nbi.dk:2135
Initializing LDAP query to grid.nbi.dk:2135
Getting LDAP query results from grid.nbi.dk:2135
Initializing LDAP connection to grid.uio.no
Initializing LDAP connection to grid.fi.uib.no
Initializing LDAP connection to fire.ii.uib.no
Initializing LDAP connection to grid.nbi.dk
Initializing LDAP connection to ns1.nordita.dk
Initializing LDAP connection to hepax1.nbi.dk
Initializing LDAP connection to lscf.nbi.dk
Initializing LDAP connection to grid.tsl.uu.se
Initializing LDAP connection to grendel.it.uu.se
Initializing LDAP connection to grid.quark.lu.se
Initializing LDAP query to grid.uio.no
Initializing LDAP query to grid.fi.uib.no
Initializing LDAP query to fire.ii.uib.no
Initializing LDAP query to grid.nbi.dk
Initializing LDAP query to ns1.nordita.dk
Initializing LDAP query to hepax1.nbi.dk
Initializing LDAP query to lscf.nbi.dk
Initializing LDAP query to grid.tsl.uu.se
Initializing LDAP query to grendel.it.uu.se
Initializing LDAP query to grid.quark.lu.se
Getting LDAP query results from grid.uio.no
Getting LDAP query results from grid.fi.uib.no
Getting LDAP query results from fire.ii.uib.no
Getting LDAP query results from grid.nbi.dk
Getting LDAP query results from ns1.nordita.dk
Getting LDAP query results from hepax1.nbi.dk
Getting LDAP query results from lscf.nbi.dk
Getting LDAP query results from grid.tsl.uu.se
Getting LDAP query results from grendel.it.uu.se
Getting LDAP query results from grid.quark.lu.se

Cluster: Oslo Grid Cluster (grid.uio.no)
Queue: default
Queue accepted as possible submission target
Cluster: Oslo Grid Cluster (grid.uio.no)
Queue: veryshort
Queue rejected because it does not match the XRSL specification
Cluster: Bergen Grid Cluster (grid.fi.uib.no)
Queue: default
Queue accepted as possible submission target
Cluster: Parallab IBM Cluster (fire.ii.uib.no)
Queue: dque
Queue rejected because user not authorized
Cluster: Copenhagen Grid Cluster (grid.nbi.dk)
Queue: long
Queue accepted as possible submission target
Cluster: Copenhagen Grid Cluster (grid.nbi.dk)
Queue: short
Queue accepted as possible submission target
Cluster: Copenhagen Nordita Cluster (ns1.nordita.dk)
Queue: p-long
Queue rejected because it does not match the XRSL specification
Cluster: Copenhagen Nordita Cluster (ns1.nordita.dk)
Queue: p-medium
Queue rejected because it does not match the XRSL specification
Cluster: Copenhagen Nordita Cluster (ns1.nordita.dk)
Queue: p-short
Queue rejected due to status: inactive
Cluster: Copenhagen Alpha Linux Machine (hepax1.nbi.dk)
Queue: long
Queue rejected due to status:
Cluster: Copenhagen Alpha Linux Machine (hepax1.nbi.dk)
Queue: short
Queue rejected due to status:
Cluster: Copenhagen LSCF Cluster (lscf.nbi.dk)
Queue: gridlong
Queue rejected due to status:
Cluster: Copenhagen LSCF Cluster (lscf.nbi.dk)
Queue: gridshort
Queue rejected due to status:
Cluster: Uppsala Grid Cluster (grid.tsl.uu.se)
Queue: default
Queue accepted as possible submission target
Cluster: Uppsala Grendel Cluster (grendel.it.uu.se)
Queue: workq
Queue accepted as possible submission target
Cluster: Lund Grid Cluster (grid.quark.lu.se)
Queue: pc
Queue accepted as possible submission target
Cluster: Lund Grid Cluster (grid.quark.lu.se)
Queue: pclong
Queue rejected because it does not match the XRSL specification

Uppsala Grendel Cluster (grendel.it.uu.se) selected
queue workq selected
Job submitted with jobid: grendel.it.uu.se:2119/jobmanager-ng/223411027195684
```

Information system

Data Management

Resource & Job Management

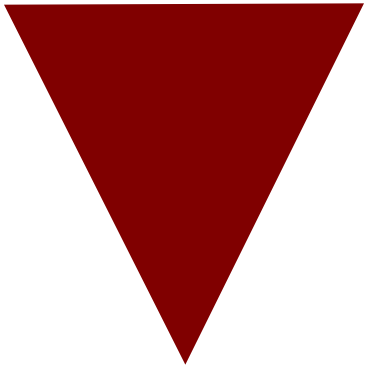
Information System

+ security

a) resource characterization / description

b) resource discovery

c) monitoring of services / resources



The nerve system of the Grid information is a critical resource on the Grid

Nordic Testbed for Wide Area Computing and Data Handling

The challenge

- large number of resources
  => scalability
- diverse heterogeneous resources
  => characterization?
- decentralized, automatic maintenance
- efficient access to dynamic data
- quality and reliability of information
  => fake information can 'kill' the Grid

Nordic Testbed for Wide Area Computing and Data Handling

challenge cont.

Grid users always want prompt access to all the information

inevitable compromise:

load on the Grid <=> up-to-dateness

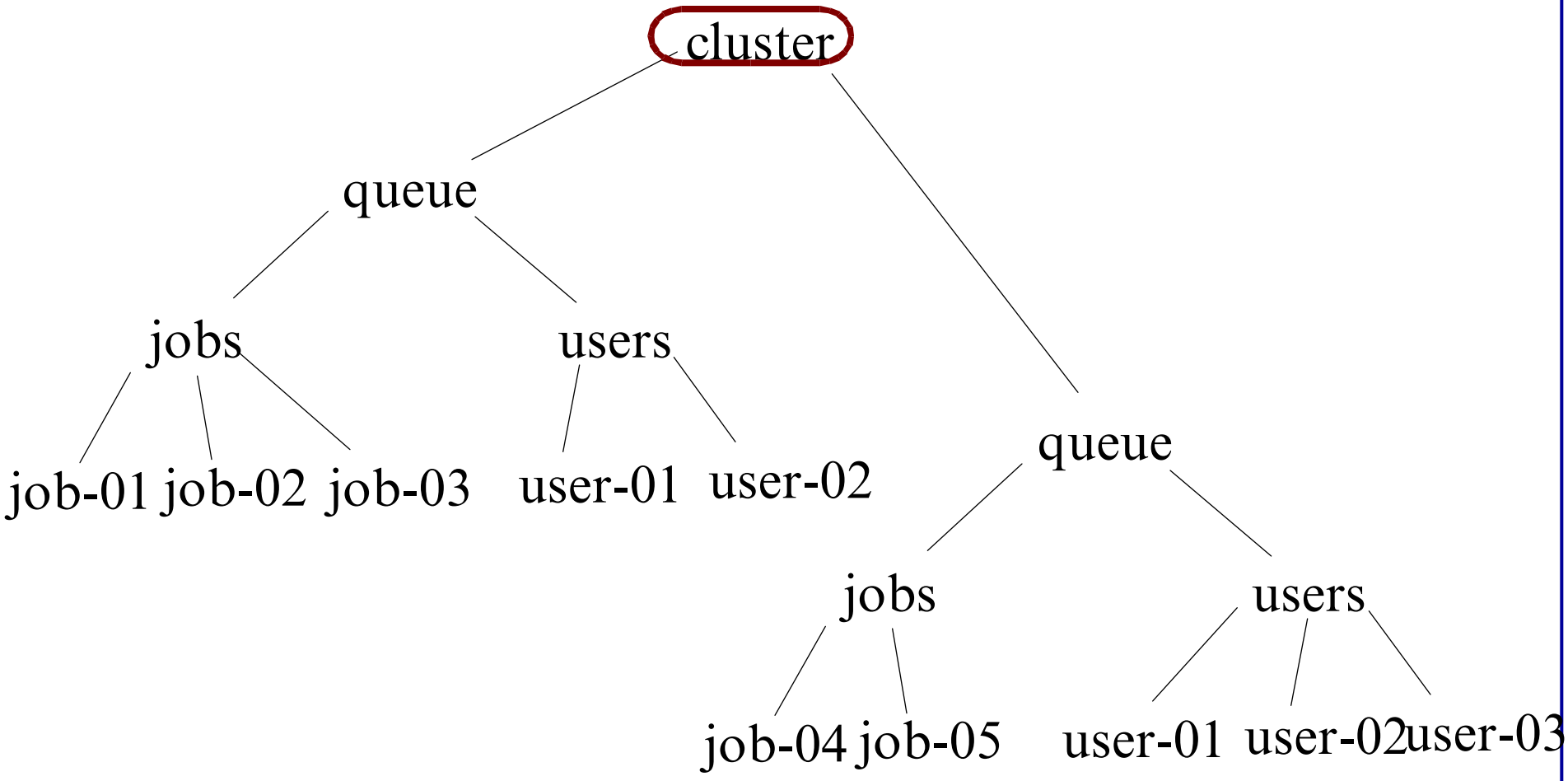
- try to avoid continuous monitoring
- generate information on demand (pull model)
- apply elaborate caching and keep track of validity of the data (ttl)
- organize “information producers” into some kind of topology (i.e. hierarchy)

The NorduGrid solution

NorduGrid Information System:

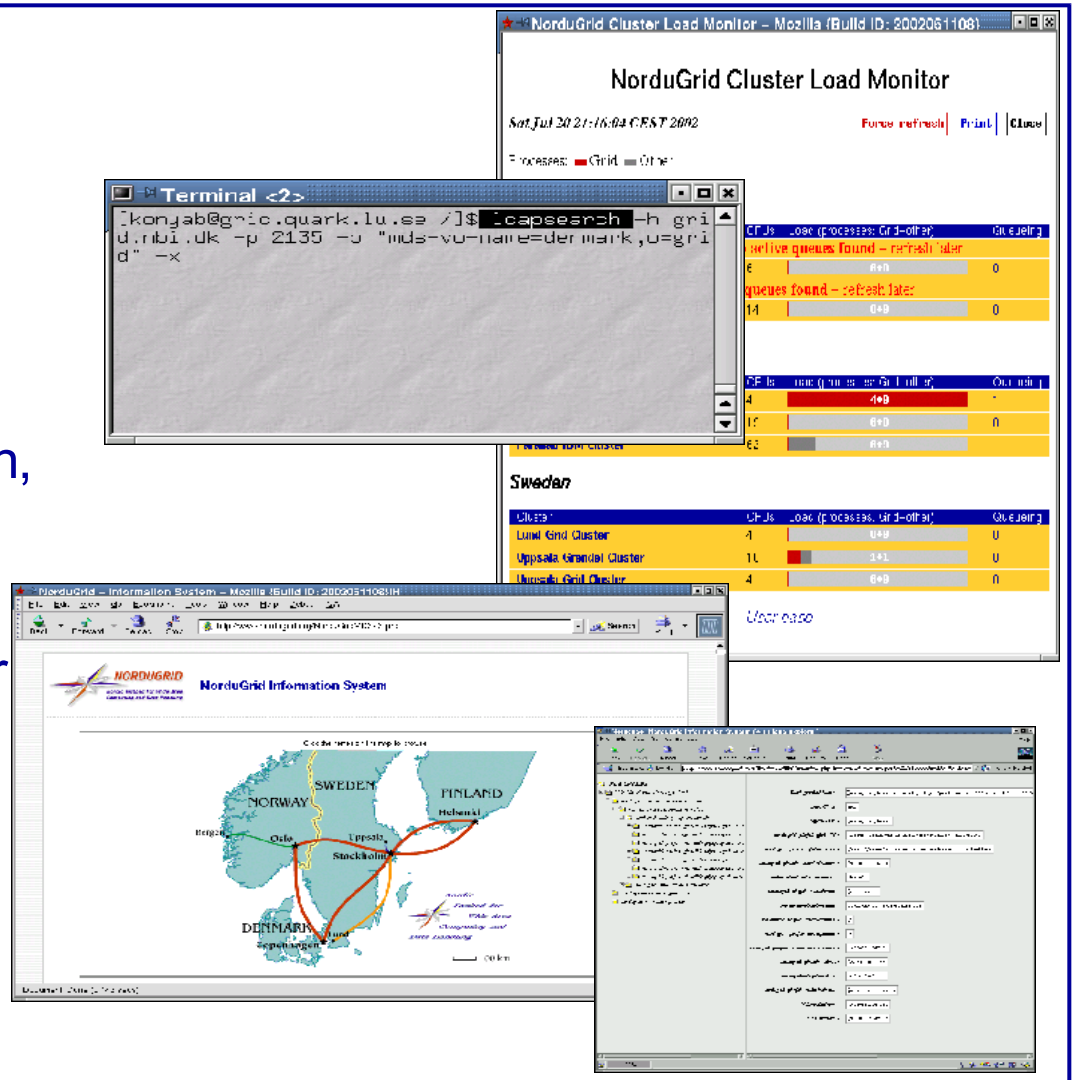
- built upon the MDS (Monitoring and Discovery Service) LDAP backends of Globus Toolkit
- the NorduGrid schema gives a natural representation of our resources
  - clusters (queues, jobs, users)
  - storage elements
  - replica catalog
- efficient providers fill the entries of the schema
- each “grid unit” runs its own (Grid Resource Information Service) GRIS
- GRISes are organized into a dynamic country-based GIIS hierarchy (Grid Index Information Service, a kind of link collection with caching)

DIT of a cluster



interfacing to the IS

- The information system speaks LDAP, easy to interface:
  - users with command line ldapsearch
  - ng-userinterface (submission, brokering, job monitoring) through LDAP C API
  - Load Monitor, MDS browser through PHP LDAP API



Nordic Testbed for Wide Area Computing and Data Handling

cluster entry

NorduGrid Cluster Details for grid.quark.lu.se [Force refresh] [Print] [Close]

Attribute	Value
Distinguished name	nordugrid-cluster-name=grid.quark.lu.se,Mds-Vo-name=local,o=grid
objectClass	Mds
	nordugrid-cluster
Front-end domain name	grid.quark.lu.se
Cluster alias	Lund Grid Cluster
Contact string	gsiftp://grid.quark.lu.se:2811/jobs
E-mail contact	grid.siteadmin@quark.lu.se
	grid.support@quark.lu.se
LRMS type	OpenPBS
LRMS version	2.3.12
LRMS details	FIFO scheduler, single job per processors
Architecture	i686
Operating system	Linux 2.4.3-20mdk
Homogeneous cluster	True
CPU type (slowest)	Pentium III (Coppermine) 1001 MHz
Memory (MB, smallest)	256
Total CPUs	4
CPU:machines	2cpu:2
Occupied CPUs	0
Queued jobs	0
Total amount of jobs	0
Local Storage Element	nordugrid-se-name=grid.quark.lu.se,Mds-Vo-name=Sweden,o=grid
Session directories area	/jobs
Unallocated disk space (MB)	28430
Grid middleware	globus-2.0-0.7ng
	nordugrid-0.2.0
Runtime environment	ATLAS-3.0.1
	ATLAS-3.2.1
	DC1-ATLAS-3.2.1
Info valid from (GMT)	20-07-2002 13:03:14
Info valid to (GMT)	20-07-2002 13:03:44

Nordic Testbed for Wide Area Computing and Data Handling

queue entry

Queue pc at grid.quark.lu.se

Force refresh | Print | Close

Attribute	Value
Distinguished name	nordugrid-pbsqueue-name=pc,nordugrid-cluster-name=grid.quark.lu.se,Mds-Vo-name=local,o=grid
objectClass	Mds
	nordugrid-pbsqueue
Queue name	pc
Queue status	active
Running jobs	3
Running Grid jobs	3
Queued jobs	1
Queued Grid jobs	1
Max. running jobs	4
Max. jobs per Unix user	3
Max. CPU time (min)	120
Default CPU time (min)	120
Scheduling policy	strict FIFO
Processors per queue	4
Info valid from (GMT)	20-07-2002 13:17:14
Info valid to (GMT)	20-07-2002 13:17:44

Nordic Testbed for Wide Area Computing and Data Handling

job entry

Job ID: gsiftp://grid.fi.uib.no:2811/jobs/9355470781464331336

Force refresh | Print | Close

Attribute	Value
Distinguished name	nordugrid-pbsjob-globalid=gsiftp://grid.fi.uib.no:2811/jobs/9355470781464331336, nordugrid-info-
objectClass	Mds
	nordugrid-pbsjob
ID	gsiftp://grid.fi.uib.no:2811/jobs/9355470781464331336
Owner	/O=Grid/O=NorduGrid/OU=uio.no/CN=Aleksandr Konstantinov
Job name	dc1.002000.simul.01101.hlt.pythia_jet_17
Job submission time (GMT)	19-07-2002 20:30:13
Execution queue	default
Execution cluster	grid.fi.uib.no
Job status	INLRMS: R
Used CPU time	1021
Used wall time	1024
Used memory (KB)	130184
Requested CPU time	2880
PBS comment	Job started on Fri Jul 19 at 22:30
Standard output file	out.txt
Standard error file	out.txt
Submission machine	129.240.86.18:4650;grid.uio.no
Info valid from (GMT)	20-07-2002 13:36:17
Info valid to (GMT)	20-07-2002 13:36:47

job status monitoring = information system query

another job entry

Job ID: gsiftp://grid.quark.lu.se:2811/jobs/18334158781110508307

Force refresh | Print | Close

Attribute	Value
Distinguished name	nordugrid-pbsjob-globalid=gsiftp://grid.quark.lu.se:2811/jobs/18334158781110508307, nordugrid-ir
objectClass	Mds
	nordugrid-pbsjob
ID	gsiftp://grid.quark.lu.se:2811/jobs/18334158781110508307
Owner	/O=Grid/O=NorduGrid/OU=quark.lu.se/CN=Balazs Konya
Job name	DC1 test at Lund
Job submission time (GMT)	19-07-2002 15:53:50
Execution queue	pc
Execution cluster	grid.quark.lu.se
Job status	FINISHED at: 20020719161437Z
Used wall time	19
Used CPU time	18
Job erase time (GMT)	20-07-2002 16:14:37
Standard output file	dc1.002000.test.NG.out
Standard error file	dc1.002000.test.NG.out
Submission machine	130.235.92.242:55972;grid.quark.lu.se
Info valid from (GMT)	20-07-2002 13:40:14
Info valid to (GMT)	20-07-2002 13:40:44

- the job entry is generated on the execution cluster
- when the job is completed and the results are retrieved the job disappears from the information system

personalized information

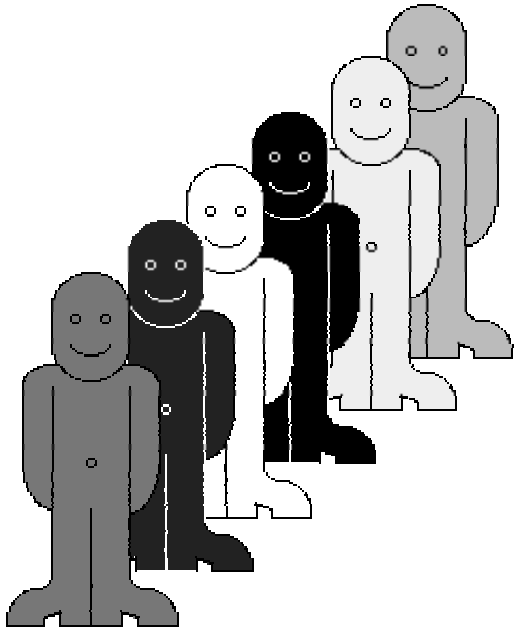
user based information is essential

on the Grid:

- users are not really interested in the total number of cpus of a cluster, but how many of those are available for them!
- number of queuing jobs are irrelevant if the submission gets immediately executed
- instead of total disk space the user's quota is interesting

nordugrid-authuser objectclass

- freecpus
- diskspace
- queuelength



Nordic Testbed for Wide Area Computing and Data Handling

user entry

```
Distinguished Name = nordugrid-authuser-name=Oxana Smirnova_14,nordugrid-i
       objectClass = Mds
       objectClass = nordugrid-authuser
nordugrid-authuser-name = Oxana Smirnova_14
nordugrid-authuser-sn = /O=Grid/O=NorduGrid/OU=quark.lu.se/CN=Oxana Smirnova
nordugrid-authuser-freecpus = 3
nordugrid-authuser-queuelength = 0
nordugrid-authuser-diskspace = 28278
Mds-validfrom = 20020720142938Z
Mds-validto = 20020720143008Z
```

XRSL

XRSL is the language in which the user formulates her job request in terms of:

- required input data
- binary, preinstalled software
- outputfiles
- resource requirements (cpu, diskspace, etc..)
- misc: email notification, debug information

RSL stands for Resource Specification Language. Introduced by Globus to communicate job requirements. NorduGrid has made some necesarry extensions: created the XRSL

Nordic Testbed for Wide Area Computing and Data Handling

XRSL cont.

The most important xrsl attributes:

inputFiles=(<file> [<location>]) ... - list of files to be transferred to the computing node from a given location

outputFiles=(<file> [<location>]) ... - list of files to be preserved after the job completion and transferred to a given location.

executables=<file1> <file2> ... - list of files to be given executable permissions.

notify=<options> <email> ... - E-mail notification on job status change.

Nordic Testbed for Wide Area Computing and Data Handling

XRSL cont.

runTimeEnvironment=<string>... - application-specific runtime environment (e.g., ATLAS-3.2.1)

middleware=<string> -required middleware (e.g., NorduGrid-0.3.0)

cluster=<string> -specific cluster request

rerun=<number> -number of attempts to re-run the job

lifeTime=<number> -maximum time for the session directory to remain on the execution node (can not override local policy)

ftpThreads=<number> -number of GridFTP threads to be used for file transfers

an example job request

```
&
(executable="my_binary.bin")
(inputFiles=
(“data12.inp”
“rc://@grid.uio.no/lc=my_files,rc=NorduGrid,dc=nordugrid,dc=org”)
(“basefile” “gsiftp://grid.quark.lu.se/nordugrid/graphics/bigdata.pxi))
(outputFiles=
(“figure.ppm”
“rc://grid.uio.no/lc=test,rc=NorduGrid,dc=nordugrid,dc=org"))
(jobName=”graphics12”)
(stdin="parameters.inp")
(stdout="stdout")
(join=yes)
(ftpThreads=6)
(middleware="NorduGrid-0.3.9")
(runtimeEnvironment=”Graphics”)
```