Nordic
Testbed for
Wide Area
Computing and
Data Handling

The NorduGrid Toolkit

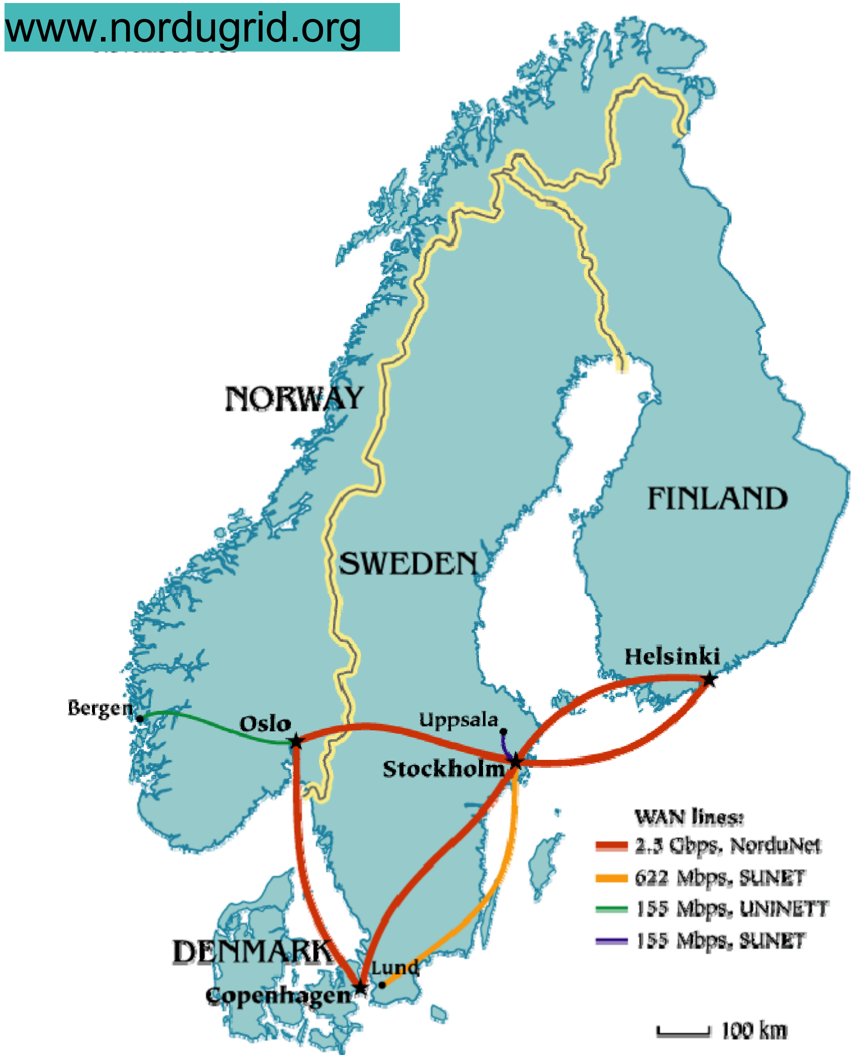
(live)

Balázs Kónya
EDG5
2nd September, 2002, Budapest

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

NorduGrid Toolkit:

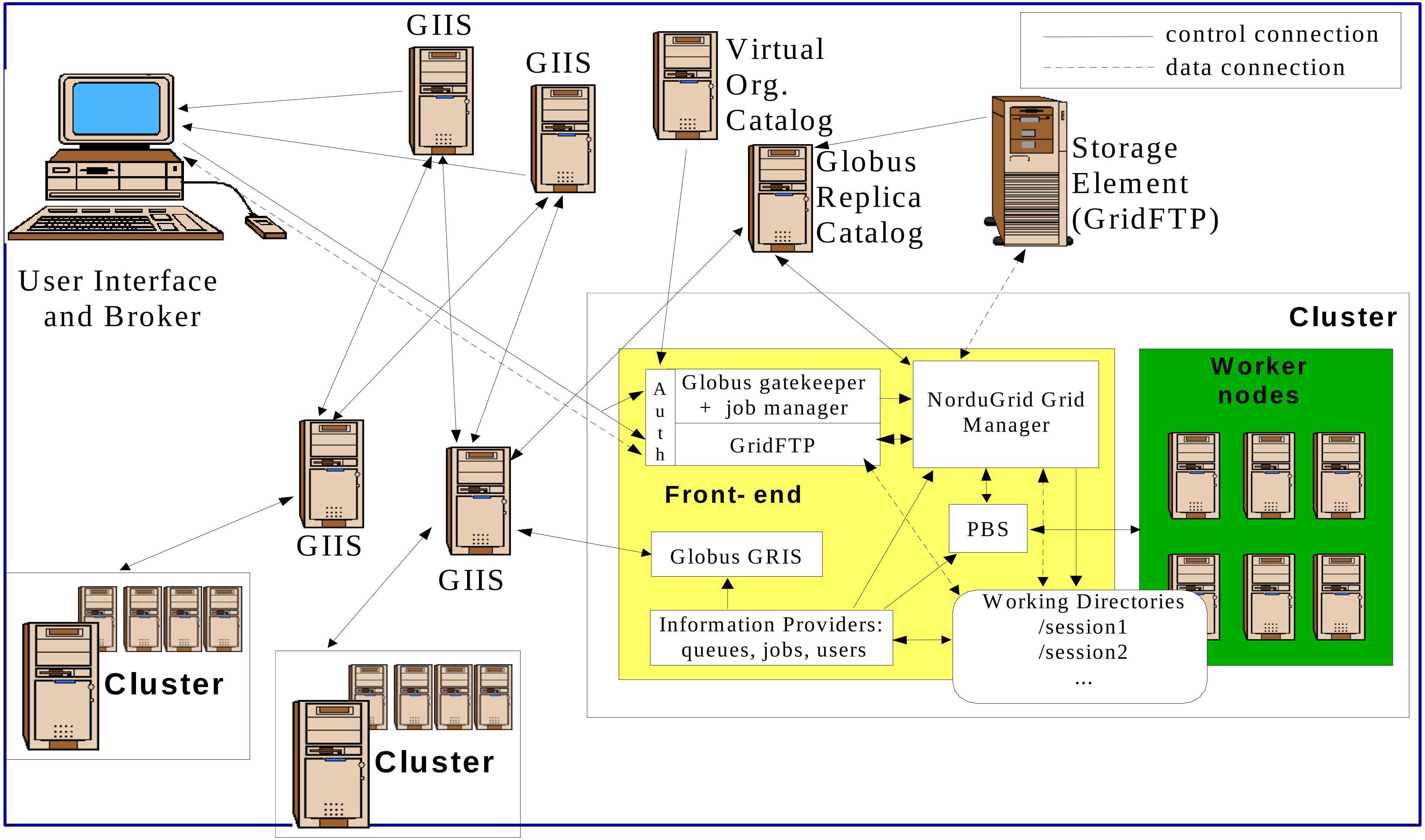
- **it is:**
  - a functional middleware solution developed by the NorduGrid project
  - implements the fundamental Grid services
  - extends the Globus Toolkit
  - replaces/obsolates some of the Globus core services
- **it is not:**
  - just a webinterface, a monitoring tool
  - an oversimplified Grid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

design considerations

- A Grid **middleware must be as simple as possible** in terms of number of
  - used protocols
  - entry points/communication channels to Grid resources
  - running Grid daemons
  - requirements imposed on participating sites
- **heterogeneous, non-dedicated** clusters, no special requirements for cluster nodes!
- The Grid is a **distributed system**, no single point of failure, no centralized services

Nordic Testbed for Wide Area Computing and Data Handling

NorduGrid architecture



the components

- **Grid Manager** (clever stage in/stage out, job management on the cluster)
- **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:**
  - storage manager
  - distributed replica manager
  - better AAA, “Grid access control”

Nordic Testbed for Wide Area Computing and Data Handling

Grid Manager

- Provide job control and data handling functionalities
- the middleware layer which sits/runs on top of the LRMS
- extends and takes over the functionality of the Globus jobmanager
- 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
  - keep results on cluster untill user downloads.

Grid Manager cont.

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

UserInterface

- **command line tools for:**

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

- **built-in brokering**

UserInterface cont.

- The UI processes user-level xRSL request and transforms to a form suitable for GM
- Performs brokering
  - analyzes information about the different clusters obtained from the MDS
  - 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 GRAM or GridFTP interface and uploads input files.
- Can be used as an MDS interface for job & cluster status

built-in brokering

1) searches through the NorduGrid Testbed for available clusters
2) loops through all the clusters and selects those queues (possible targets) where:
   - the user is authorized to run
   - the requested software (RuntimeEnvironment) is available
   - the cluster & queue parameters match the job requests
3) selects a job destination from the matching targets
   a) randomly selects among the free resources (where user-freecpus >0)
   b) in case there are no free matching resources some of the “load” attributes (i.e. user-queuelength) are taken into account

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
```

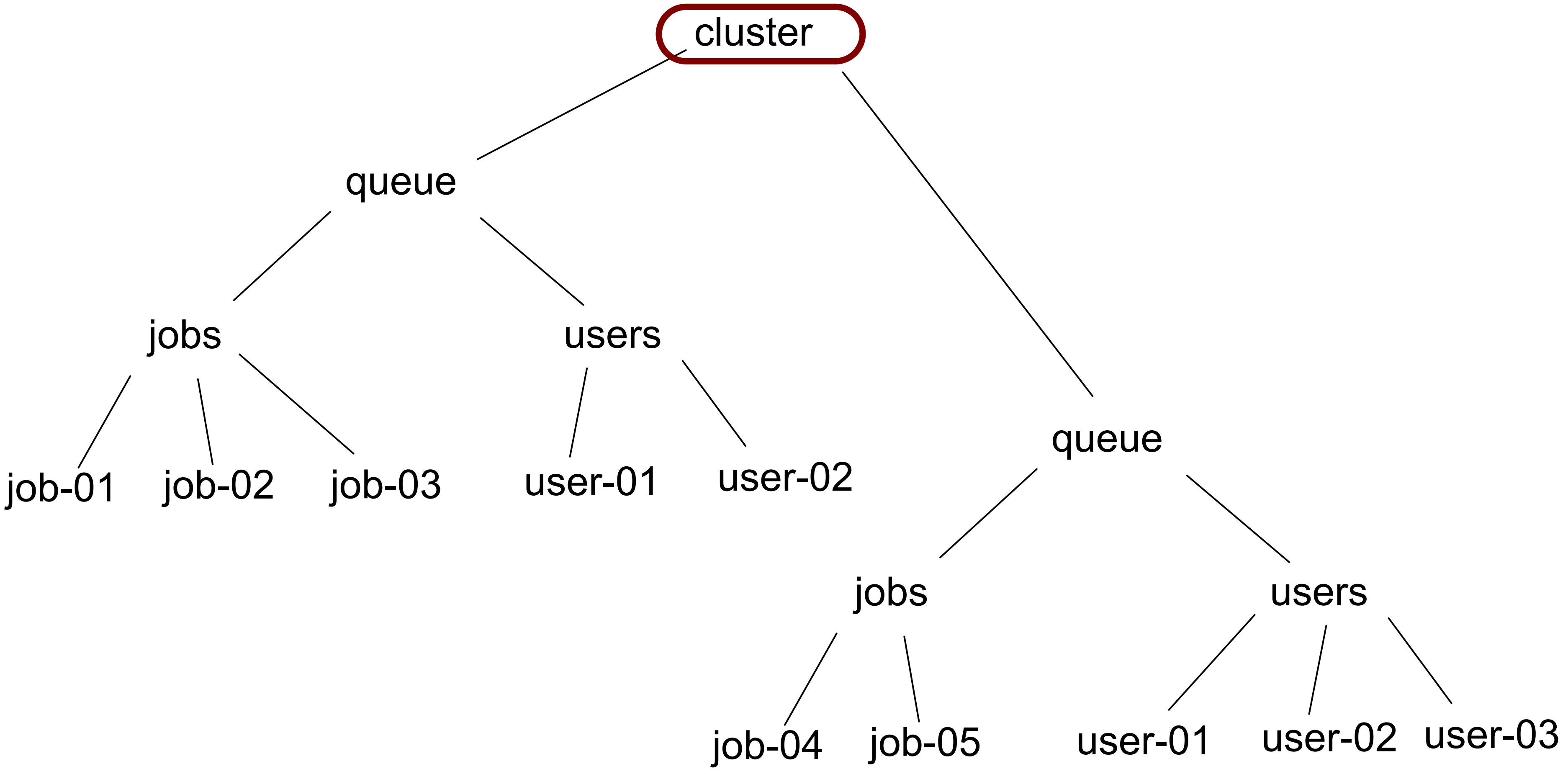
```
Cluster: Parallab IBM Cluster (fire.ii.uib.no)
Queue: dque
Queue rejected because user not authorized
Cluster: Copenhagen Grid Cluster (grid.nbi.dk)
Queue: long
Queue accepted as possible submission target
Cluster: Copenhagen Grid Cluster (grid.nbi.dk)
Queue: short
Queue accepted as possible submission 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

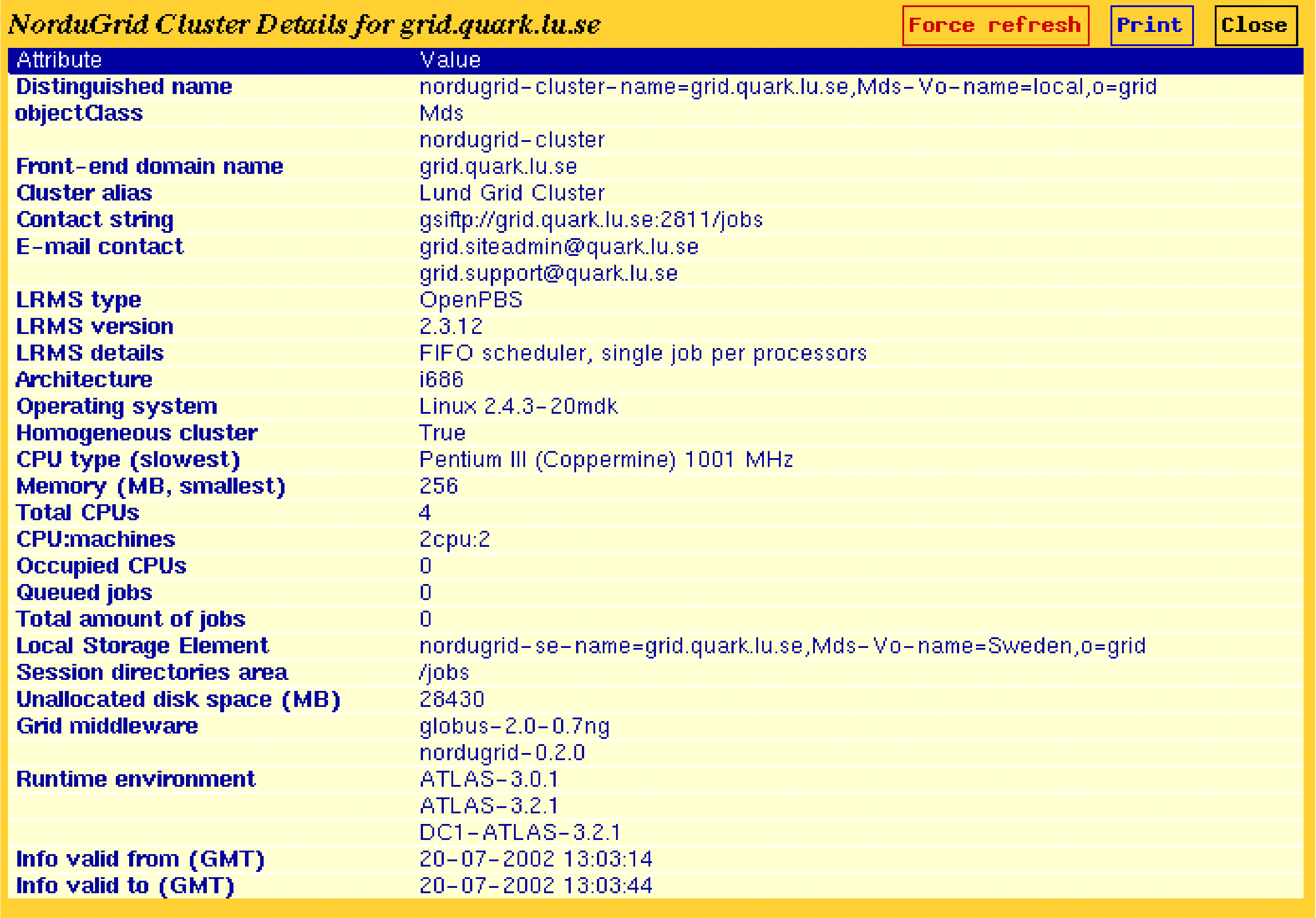
NorduGrid Information System:

- built upon the MDS 2.2 LDAP backends
- 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 **GRIS**
- GRISes are organized into a dynamic **country-based GIIS hierarchy**

DIT of a cluster



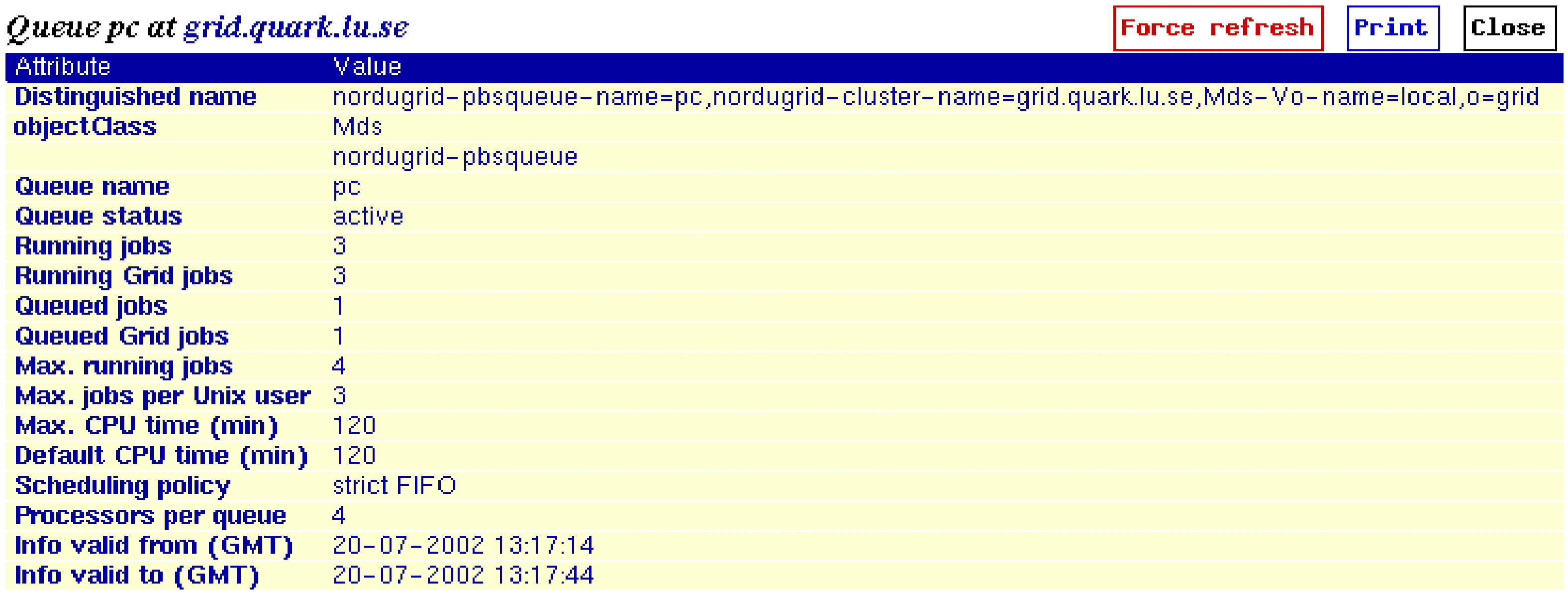
cluster entry



NorduGrid Cluster Details for grid.quark.lu.se

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

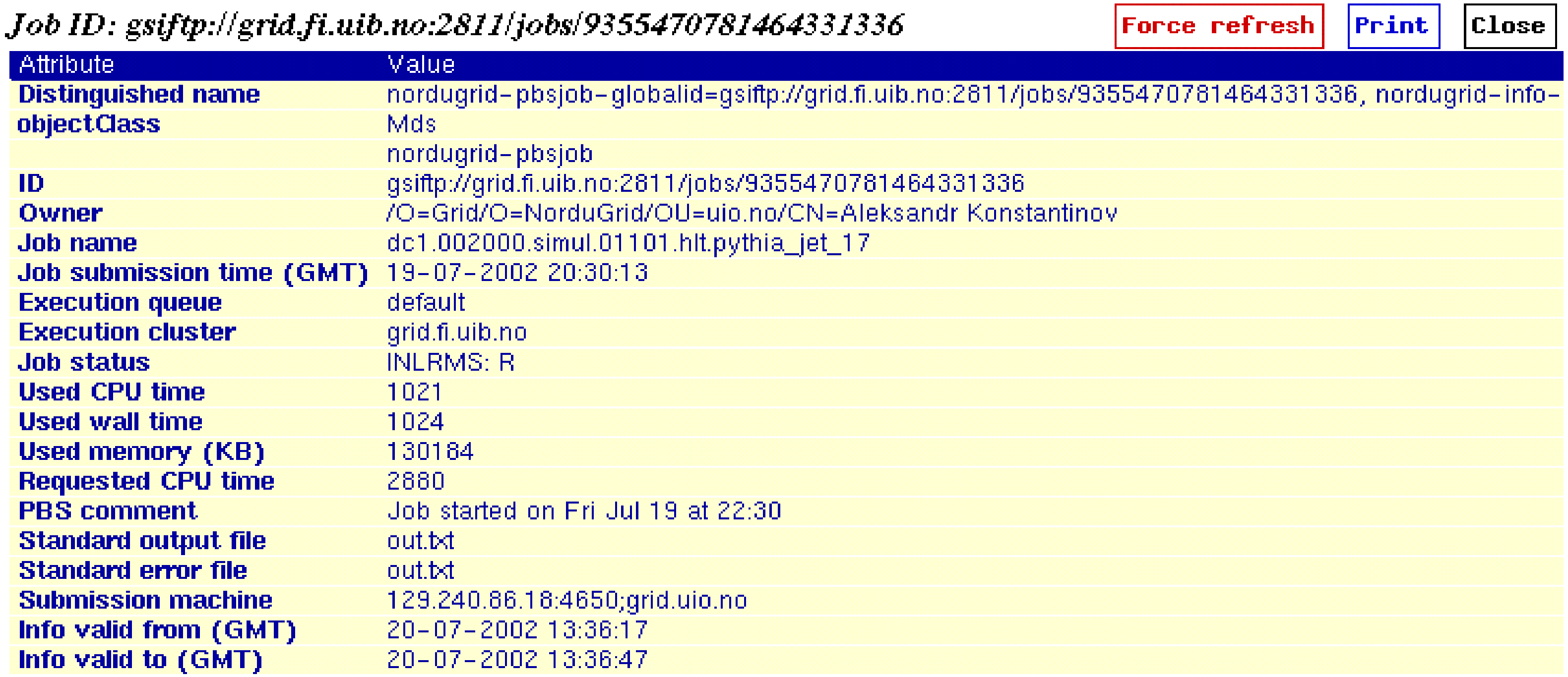
queue entry



Queue pc at grid.quark.lu.se

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

job entry

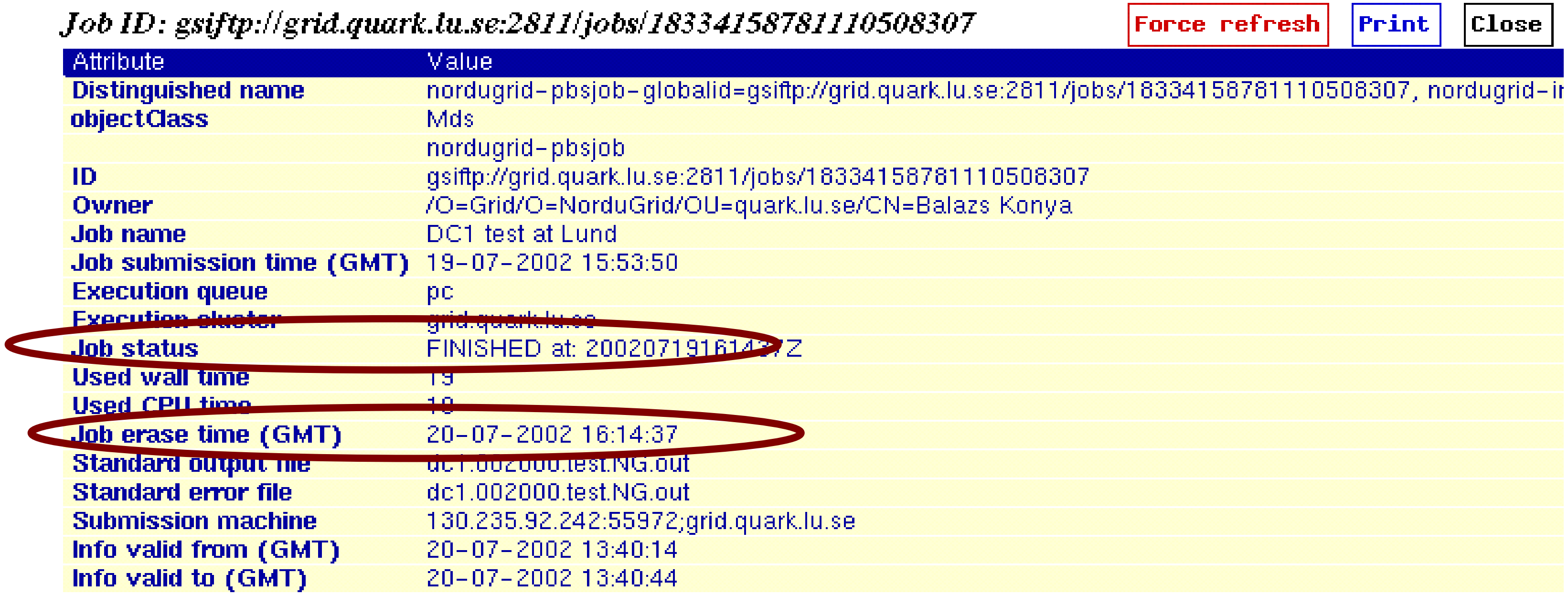


Job ID: gsiftp://grid.fi.uib.no:2811/jobs/93554707814643313336

Attribute	Value
Distinguished name	nordugrid-pbsjob-globalid=gsiftp://grid.fi.uib.no:2811/jobs/93554707814643313336, nordugrid-info-
objectClass	Mds
	nordugrid-pbsjob
ID	gsiftp://grid.fi.uib.no:2811/jobs/93554707814643313336
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

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

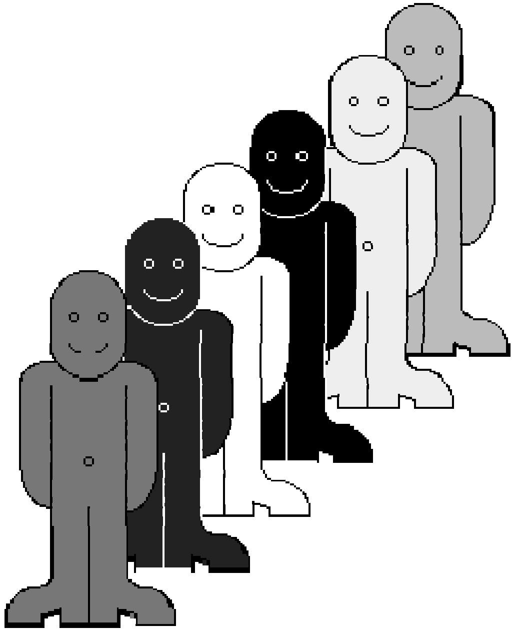
personalized information

user based information is **essential** on the Grid:

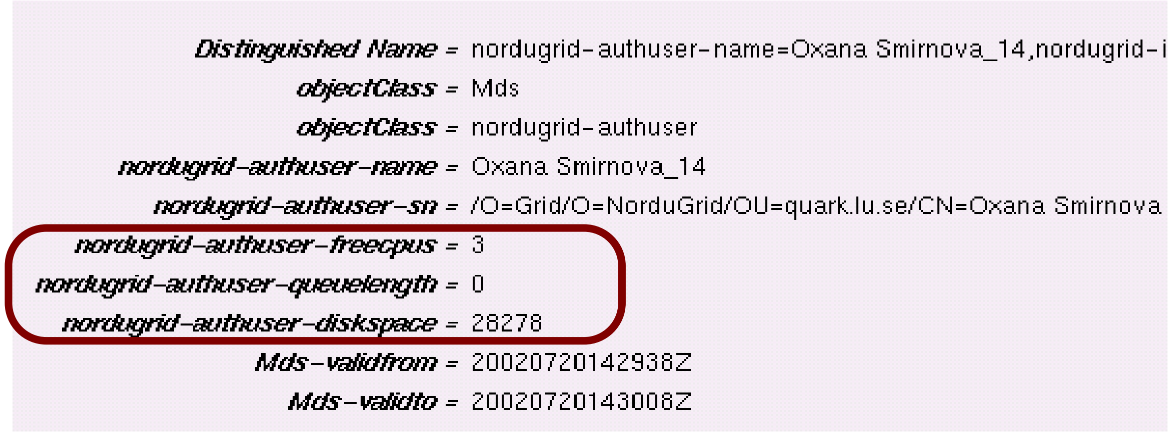
- users are not really interested in the total number of cpus of a cluster, but how many of those are available for them!
- number of queuing jobs are irrelevant if the submission gets immediately executed
- instead of total disk space the user's quota is interesting

nordugrid-authuser objectclass

- freecpus
- diskspace
- queuelength



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
```

Extended *RSL*

RSL stands for Resource Specification Language. Introduced by Globus to communicate job requirements to the Global Resource Allocation Manager (GRAM):

- **Allows basic logical expressions**
- **Set of attributes is expandable**
- **Unknown attributes are passed through.**
  - **Allows different parts to be processed at different levels.**
  - **Can be used to assist in writing brokers or filters which refine an RSL specification**

XRSL (new attributes)

To support additional features new attributes introduced. The most important are

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.

XRSL (new attributes)

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)

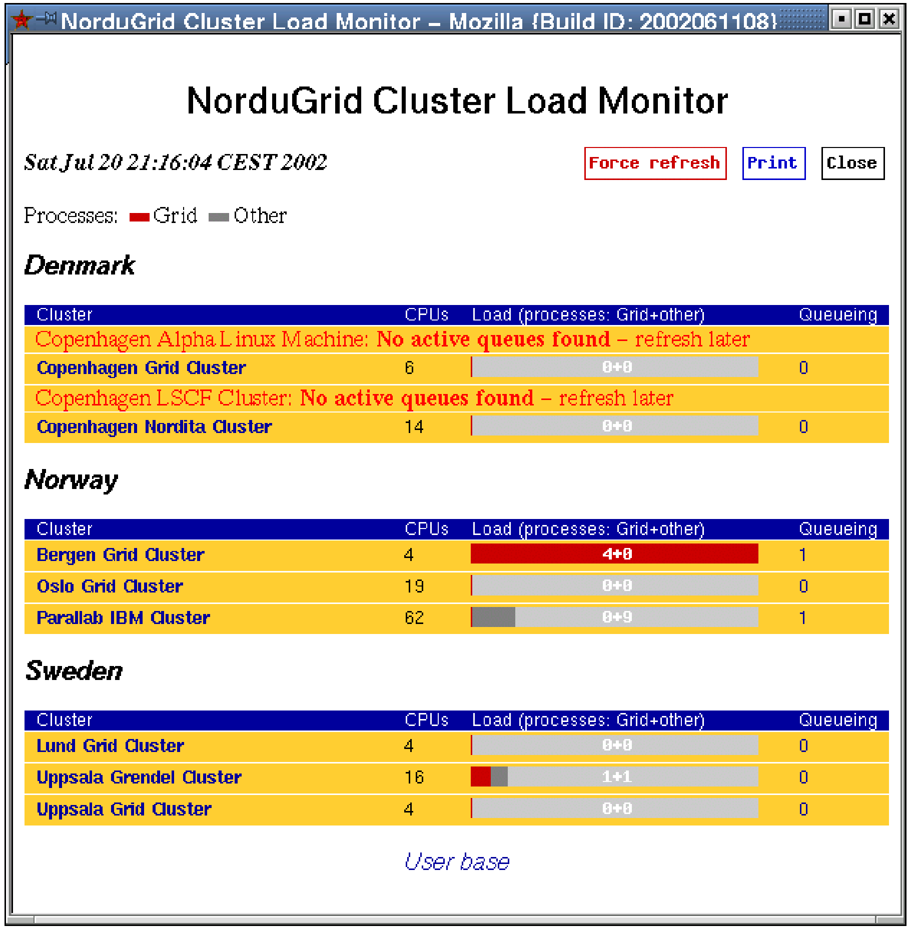
ftpThreads=<number> -number of GridFTP threads to be used for file transfers

an example xrsl

```
&
(executable="my_binary.bin")
(*inputFiles=(data.inp “gsiftp://se.nordugrid.org/disk/1002.dat”)*)
(outputFiles=(figure.ppm
                                        “rc://grid.uio.no/lc=test,rc=NorduGrid,dc=nordugrid,dc=org"))
(jobName=mandelbrot)
(stdin="parameters.inp")
(stdout="stdout")
(join=yes)
(ftpThreads=6)
(middleware="NorduGrid-0.3.4")
(*runtimeEnvironment=”Graphics”*)
```

LoadMonitor

- thanks to Oxana we have a very nice monitoring interface (through LDAP/PHP) to the MDS
- dynamic view of the
  - TestBed status
  - user activity
  - job status information
  - etc...



NorduGrid Cluster Load Monitor – Mozilla {Build ID: 2002061108}

NorduGrid Cluster Load Monitor

Sat Jul 20 21:16:04 CEST 2002

Force refresh
Print
Close

Processes: 🟥 Grid ▬ Other

Denmark

Cluster	CPUs	Load (processes: Grid+other)	Queueing
Copenhagen Alpha Linux Machine: **No active queues found** – refresh later			
Copenhagen Grid Cluster	6	0+0	0
Copenhagen LSCF Cluster: **No active queues found** – refresh later			
Copenhagen Nordita Cluster	14	0+0	0

Norway

Cluster	CPUs	Load (processes: Grid+other)	Queueing
Bergen Grid Cluster	4	4+0	1
Oslo Grid Cluster	19	0+0	0
Parallab IBM Cluster	62	0+9	1

Sweden

Cluster	CPUs	Load (processes: Grid+other)	Queueing
Lund Grid Cluster	4	0+0	0
Uppsala Grendel Cluster	16	1+1	0
Uppsala Grid Cluster	4	0+0	0

User base

Nordic Testbed for Wide Area Computing and Data Handling

Conclusions

- The Globus toolkit *alone* is not sufficient for a functional TestBed, but provides a solid development base.
- The NorduGrid Toolkit extends the Globus Toolkit and provides a working environment for Grid computing.
  - gridmanager
  - xrsl
  - userinterface (built in broker)
  - information model/system
  - cluster monitor
- The Toolkit is under continous testing in a production quality TestBed
- A lot of things to do:
  - interactive access, runtime data handling, distributed replica catalog, accounting, parallel jobs, better support for different LRMS, improved brokering algorithms, etc...

further information

- documentation:
  - papers on GM, UI, XRSL, infosys
  - www.nordugrid.org/documents
- software repository:
  - www.nordugrid.org/software
- mailing lists:
  - nordugrid-discuss, nordugrid-support

The NorduGrid core team :

Mattias Ellert
Aleksandr Konstantinov
Balázs Kónya
Jakob Langgaard Nielsen
Oxana Smirnova
Anders Wäänänen