The NorduGrid Information System

Balázs Kónya
4th NorduGrid Workshop
11-13 November, 2002, Uppsala
summary

'anno the 3\textsuperscript{rd} NG Workshop, May 2002, Helsinki'

- NorduGrid TestBed is running on a distributed dynamic information system:
  - NorduGrid Information Model
  - efficient providers
  - the resources are organized into a country-based hierarchy
  - dynamic site registration
  - the brokering is done by using the dynamic attributes
  - user-friendly interfaces
'anno the 3rd NG Workshop, May 2002, Helsinki''

- more sites to join
  - real-life scalability tests
  - more complex hierarchy, new topology?
  - performance tests with real-life applications
  - feedback from users/siteadmins

- increased stability of the Globus software with MDS 2.2

- expected new functionality from the Globus (invitation, dynamic access control to the information)

- support for more LRMS (GridEngine, LSF, ...)

- Common Grid Information Model from the GGF
'as of the 4\textsuperscript{rd} NG Workshop, November 2002, Uppsala'

- we have several new sites joined the TestBed
  - small scale test clusters
  - big production clusters: Umea, Linköping, Parallab, more to come: Monolith
  - NorduGrid is the largest operational dynamic TestBed of the world

- Performance, scalability, feedback
  - successful Atlas DC1 production runs
  - the providers scale “smoothly” both on small and large clusters
  - experimenting with different topologies
  - user/sysadmin feedback:
    - support for non-dedicated /dedicated clusters
    - we plan to implement some support for parallel jobs on Wednesday
• stability problems with the Globus software:
  ● Globus does not TEST the software
  ● we are the only serious users -> all the problems pop up on the NorduGrid
  ● most of the severe bugs can not be discovered on small-size unused testbeds -> NorduGrid TestBed plays an essential role
  ● we have contributed several bugfixes (not really getting acknowledged)

• new functionality of the Globus MDS 2.2
  ● no dynamic access control, no invitation
  ● authenticated (secure) access is unusable (huge memory leaks)

• Support for more LRMS: PDC is porting to the Easy scheduler

• GGF standard for the Grid Information Model: very far in the future (see the next slide: “Tale of the common schema”)
Once upon a time there was the MDS 2.0 “the son of the old Globus”

- grown up among his old friends the “MPPs” & “traditional supercomputers” he lived happily on the Grid
- then appeared the clusters and the boy found itself completely lost

Meanwhile a new settler (called the NorduGrid) arrived to the Grid and wanted learn the language of the clusters

- unfortunately nobody knew the language of the clusters, even the big Prophets from the Forum had nothing in their Holy Books
- so the NorduGrid started to learn the language of the clusters on his own
  - built a new village for the clusters on the Grid (he called it the TestBed)
  - as living among the clusters in the new village he slowly understood what are the important words, expressions of the clusters
- then more and more clusters moved to the new NorduGrid village and during the evenings the happy clusters used to tell horror stories about villages where there was nothing else but CHAOS
Then NorduGrid decided to share his dictionary (called schema) with the other settlers of the Grid, however it was too late:

- at the same time the old Globus had hired a self-appointed Wizard called himself “Glue” to teach his son MDS the language of clusters
- it turned out that the Wizard had never talked to any cluster before, had not lived among the clusters, never built a village for them
- but he succeeded to convince the new settlers of the Grid that he is “the authentic one” who will decode the language of the clusters
- when NorduGrid appeared at the Forum, and demonstrated that he could talk to the clusters nobody was really interested, it was already too late, everybody wanted to listen to the Wizard

The Wizard promised everything to the settlers and the Prophets of the Forum started to believe in the words of the Wizard: The Prophets thought they had to follow the Wizard since he was the one chosen by the old Globus.
Time was passing by and the new “Glue dictionary” was still in its very early stage:

- Once the boy MDS went to a village of clusters but the clusters did not understand him.

- Meanwhile NorduGrid started to correct the enormous mistakes of the Glue dictionary, posted several corrections to the Forum, as a result the Glue dictionary looked more and more similar to the NorduGrid language.

- The Wizard traveled around the Grid and kept telling to everybody that the Glue would be the real language, he never mentioned the fact that a settler in the North in his TestBed already had been talking the language of the clusters for a long time.

- Later some of the settlers heard about NorduGrid and his village. These hypnotized settlers angrily questioned NorduGrid why he didn't change his language and started to learn the Glue.
Time passed by on the Grid and

- the peaceful village of NorduGrid had grown to the largest settlement for clusters
- there is no village where clusters understand or use the Glue language of the Wizard.
- the hypnotized settlers still keep asking NorduGrid why he does not drop his language and order the clusters in his village to talk the Glue
- as NorduGrid understands more and more the clusters, his dictionary is getting better and better
- NorduGrid does not understand the reason why the settlers don't take his dictionary and start improving it instead of requesting him to speak the limited Glue
- MDS, the son of the old Globus, still has problems with talking to the clusters
- The Wizard is still there and loudly praises the unused Glue language
- meantime the Prophets of the Forum being influenced by the spirit of the “Big Blue” decided that they need a new language, called CIM
that was just a tail...

the story has nothing to do with the reality and especially has nothing to do with the following Grid players:

- NorduGrid information system:

- MDS, the son of the “old Globus”:
  http://www.globus.org/mds

- Glue schema project:
  http://www.hicb.org/glue/glue-schema/schema.htm

- The Global Grid Forum (GGF):
  http://www.gridforum.org

- CIM group of the GGF:
  http://www.isi.edu/~flon/cgs-wg/index.htm

- some of the settlers: EDG, GriPhyn, Datatag, etc..
what is under the hood?
what is a grid information system?

representation of information

technical implementation (LDAP, …)
Grid Information System

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.
Why is it so complicated?

- 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
Grid users always want **prompt** access to all the information.

**inevitable compromise:**

- load on the Grid \(\leftrightarrow\) 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)
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**
  - **queue**
    - **jobs**
      - job-01
      - job-02
      - job-03
    - **users**
      - user-01
      - user-02
  - **queue**
    - **jobs**
      - job-04
      - job-05
    - **users**
      - user-01
      - user-02
      - user-03
### NorduGrid Cluster Details for grid.quark.lu.se

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinguished name</td>
<td>no:ugrid-cluster-name=grid.quark.lu.se,Mas-Vo-name=local,0=grid</td>
</tr>
<tr>
<td>nginx Class</td>
<td>Mas</td>
</tr>
<tr>
<td>Front-end domain name</td>
<td>gtri:quark.lu.se</td>
</tr>
<tr>
<td>Cluster alias</td>
<td>Lu-d Grid Cluster</td>
</tr>
<tr>
<td>Contact string</td>
<td><a href="mailto:gtri@quark.lu.se">gtri@quark.lu.se</a>; <a href="mailto:gri@grid.quark.lu.se">gri@grid.quark.lu.se</a>; <a href="mailto:gri@quark.lu.se">gri@quark.lu.se</a></td>
</tr>
<tr>
<td>E-mail contact</td>
<td><a href="mailto:gtri@quark.lu.se">gtri@quark.lu.se</a>; <a href="mailto:gri@grid.quark.lu.se">gri@grid.quark.lu.se</a>; <a href="mailto:gri@quark.lu.se">gri@quark.lu.se</a></td>
</tr>
<tr>
<td>LRMS type</td>
<td>OpenPBS</td>
</tr>
<tr>
<td>LRMS version</td>
<td>v.3.8</td>
</tr>
<tr>
<td>LRMS details</td>
<td>FIFC scheduler, single job per processor</td>
</tr>
<tr>
<td>Architecture</td>
<td>i866</td>
</tr>
<tr>
<td>Operating system</td>
<td>Linux 2.4.1-20mdk</td>
</tr>
<tr>
<td>Homogeneous cluster</td>
<td>n1=</td>
</tr>
<tr>
<td>CPU type (slowest)</td>
<td>Pentium II (Coppermine) 601 MHz</td>
</tr>
<tr>
<td>Memory (MB, smallest)</td>
<td>252</td>
</tr>
<tr>
<td>Total CPUs</td>
<td>4</td>
</tr>
<tr>
<td>CPU Machines</td>
<td>4</td>
</tr>
<tr>
<td>Occupied CPUs</td>
<td>0</td>
</tr>
<tr>
<td>Queued jobs</td>
<td>0</td>
</tr>
<tr>
<td>Total amount of jobs</td>
<td>0</td>
</tr>
<tr>
<td>Local Storage Element</td>
<td>no:ugrid-se-name=grid.quark.lu.se,Mas-Vo-name=local,0=grid</td>
</tr>
<tr>
<td>Session directories area</td>
<td>/jobs</td>
</tr>
<tr>
<td>Unallocated disk space (MB)</td>
<td>28130</td>
</tr>
<tr>
<td>Grid middleware</td>
<td>glibus-2.0-1.7n</td>
</tr>
<tr>
<td>Runtime environment</td>
<td>ATLAS-3.1; ATLAS-3.2; CC1-ATLAS-3.2;</td>
</tr>
<tr>
<td>Intrust valid from (GMT)</td>
<td>20/12/2012 03:34</td>
</tr>
<tr>
<td>Intrust valid to (GMT)</td>
<td>20/12/2012 03:34</td>
</tr>
</tbody>
</table>
DIT of a cluster

cluster

queue

jobs

job-01

job-02

job-03

users

user-01

user-02

queue

jobs

job-04

job-05

users

user-01

user-02

user-03
### Queue Entry

#### Attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinguished name</td>
<td>ncray.d-pbsquark-nam-pc::undu.grid-luster-nam=grid-quark.nu,Ms=Vv-nam=lu-al,v-grid</td>
</tr>
<tr>
<td>objectClass</td>
<td>Mds</td>
</tr>
<tr>
<td>Queue name</td>
<td>pc</td>
</tr>
<tr>
<td>Queue status</td>
<td>active</td>
</tr>
<tr>
<td>Running jobs</td>
<td>0</td>
</tr>
<tr>
<td>Running Grid jobs</td>
<td>3</td>
</tr>
<tr>
<td>Queued jobs</td>
<td></td>
</tr>
<tr>
<td>Queue Grid jobs</td>
<td></td>
</tr>
<tr>
<td>Max. running jobs</td>
<td>4</td>
</tr>
<tr>
<td>Max. jobs per Unix user</td>
<td>4</td>
</tr>
<tr>
<td>Max. CPU time (min)</td>
<td>25</td>
</tr>
<tr>
<td>Default CPU time (min)</td>
<td>25</td>
</tr>
<tr>
<td>Scheduling policy</td>
<td>strict FIFO</td>
</tr>
<tr>
<td>Processes per queue</td>
<td>4</td>
</tr>
<tr>
<td>Info valid from (GMT)</td>
<td>20 17 2002 13:7:14</td>
</tr>
<tr>
<td>Info valid to (GMT)</td>
<td>20-07-2002 13:7:44</td>
</tr>
</tbody>
</table>

---

12/17/02  balazs.konya@quark.lu.se
DIT of a cluster

- cluster
  - queue
    - jobs
      - job-01
      - job-02
      - job-03
    - users
      - user-01
      - user-02
job entry

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinguished name</td>
<td>ncrdug.d-pbso-glob-ali:grif.fui.no 2011/jobs/935547081464331336</td>
</tr>
<tr>
<td>objectClass</td>
<td>Mts</td>
</tr>
<tr>
<td>ID</td>
<td>grif.fui.no 2011/jobs/935547081464331336</td>
</tr>
<tr>
<td>Owner</td>
<td>/C=GridC=Kordu@Grid/OU=io/OU=Ch=Aleksi@Konstantinova</td>
</tr>
<tr>
<td>Job name</td>
<td>d-11_111111111111 http://net1_1/</td>
</tr>
<tr>
<td>Job submission time (GMT)</td>
<td>16-07-2002 22:30:13</td>
</tr>
<tr>
<td>Execution queue</td>
<td>default</td>
</tr>
<tr>
<td>Execution cluster</td>
<td>q.fui.no</td>
</tr>
<tr>
<td>Job status</td>
<td>running</td>
</tr>
<tr>
<td>Used CPU time</td>
<td>122</td>
</tr>
<tr>
<td>Used wall time</td>
<td>1:24</td>
</tr>
<tr>
<td>Used memory (KB)</td>
<td>13040</td>
</tr>
<tr>
<td>Requested CPU time</td>
<td>4-111</td>
</tr>
<tr>
<td>PBS comment</td>
<td>Job started on Fri Jul 16 at 22:30</td>
</tr>
<tr>
<td>Standard output file</td>
<td>u.txt</td>
</tr>
<tr>
<td>Standard error file</td>
<td>o.txt</td>
</tr>
<tr>
<td>Submission machine</td>
<td>1:19:4114:4114:grif.fui.no</td>
</tr>
<tr>
<td>Info valid from (GMT)</td>
<td>16-07-2002 22:36:17</td>
</tr>
<tr>
<td>Info valid to (GMT)</td>
<td>16-07-2002 22:36:17</td>
</tr>
</tbody>
</table>

**job status monitoring = information system query**
another job entry

- 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
DIT of a cluster

- Cluster
  - Queue
    - Jobs
      - job-01
      - job-02
      - job-03
    - Users
      - user-01
      - user-02
       - Jobs
         - job-04
         - job-05
       - Users
         - user-01
         - user-02
         - user-03

balazs.konya@quark.lu.se
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
Distinguished Name: cn=Csaba Sirmay, ou=csaba.sirmay, o=csaba.sirmay, c=hu

objectClass: mdm

objectClass: ndugic-authuser

ndugic-authuser-name: Csaba Sirmay

ndugic-authuser-sn = Csaba Sirmay

ndugic-authuser-tranupus = 3

ndugic-authuser-passwd = n

ndugic-authuser-diskspace = 20273

Mds valldrom = 2002072014293EZ

Mds-valldrom = 20020720143000Z
available information:
SE, RC (preliminary)

- **Storage Element**
  - **se-baseurl**
    gsiftp://bambi.quark.lu.se:2811/gamma/scratch
  - **se-freespace**
  - **se-authuser**

- **Replica Catalog**
  - **rc-baseurl**
    ldap://grid.uio.no:389/rc=NorduGrid,dc=nordugrid,dc=org
  - **rc-authuser**

*these objectclasses are 'under construction' they will gain real importance with the NorduGrid Storage Manager*
Hierarchy of GRISes/GIISeS
Security aspects

- MDS inherits OpenLDAP Version 3 security infrastructure
  - SASL (Simple Authentication and Security Layer) with GSS-API
  - secure authentication and message protection
  - GSI-GSSAPI is the Globus 'security layer' for the MDS
- In principle, user certificates can be used for accessing 'sensitive' or 'protected' information
- However:
  - only static ACL's (LDAP rules for who can access what)
  - the security is 'lost' between GRIS -- GIIS queries, queries initiated by GIISes are using the identity of the GIIS instead of the original user
  - not much of use at the moment, severe memory leaks in the software,...
• 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 (thanks to Oxana)
summary

oh not again ;) see the previous slides “present”

acknowledgment:

• NorduGrid core team
• resource providers: NSC, Parallab, PDC, HPC2N, Copenhagen, Finland
• Nordunet