The NorduGrid production Grid infrastructure

GRID2003, Phoenix, November 17, 2003

Balázs Kónya, NorduGrid Team, Lund University, Sweden
outline

- NorduGrid from a historical perspective
- What is it and what is it not?
- Middleware: the engine under the hood
- Users & Applications
- Summary
Grids are everywhere...

- Europe is flooded by Grid deployment projects
- [...] more than 6,000 grids have been deployed worldwide (Sun)
- If by deploying a scheduler on my local network I create a “Cluster Grid”, doesn't my NFS deployment over the same network provide me with a “Storage Grid?” [...] Is there any computer system that isn't a Grid? (Ian Foster)
Where are the production Grids?

- **A Grid**
  - is a decentralized system spanning multiple administrative domains, where both the set of users and resources vary dynamically
  - provides uniform access to heterogeneous systems
  - provides a flexible mechanism for locating and selecting resources based on the user's criteria

- **A production Grid**
  - Can provide a well-defined set of services with **reliable quality**
  - Serves real users and applications
NorduGrid, the Nordic Production Grid Infrastructure

- It is called NorduGrid* or Nordic Grid, because it connects the resources of Scandinavia and Finland in North Europe. “Nordu” comes from the word NorduNet which is the Nordic Internet highway to research and education in Denmark, Finland, Iceland, Norway and Sweden, and provides the Nordic backbone to the Global Information Society.
- It is a Production Grid available 24/7 since July 2002 to a growing set of users.
- It is an evolving Infrastructure with increasing number of resources and improving grid services.
- It is a Grid OPEN for participation.

*The name of the Grid is expected to be changed.

balazs.konya@hep.lu.se
www.nordugrid.org
A short history of NorduGrid

The beginnings: “Grid Book”, Globus jumps into the middle of interests
2000 September: GT-1.1.4 released
2001 January: EDG started up
2001 November: GT-2.0 was announced (very much alpha/beta software)
2002 February: OGSA Initiative
2002 April: long-awaited GT-2.0 was delivered with GRAM-1.5, MDS-2.0

2001 June: First meeting of the NorduGrid developers
2001 September: Grid Testbed with GT-1.1.4, switching to GT-2.0 pre-alpha, evaluating EDG
2002 February: decision to develop an alternative middleware by making use of Globus libraries. NorduGrid design, architecture, philosophy
2002 May: 3rd NorduGrid Workshop, Helsinki

2003-11-17 balazs.konya@hep.lu.se
www.nordugrid.org
A short history of NorduGrid

2002 July: GGF5, Edinburgh, OGSA partnership, GT-3 plans
2002 August: Globus-NorduGrid meeting in Copenhagen
2002 October: GT-2.2 released with GRAM-1.6, MDS-2.2
2003 January: GT-3 alpha released
2003 June: GT-3.0 is out

2002 May: First Atlas validation job completed
2002 July: The testbed transforms into a production Grid, the middleware is deployed throughout Scandinavian centers, the HEP group starts running their Atlas DataChallenge on the Grid
2002 July: NorduGrid is presented in one of the GGF sessions, application-driven development continues
2003 July: HEP group completes the 3rd phase of their DataChallenge

balazs.konya@hep.lu.se
www.nordugrid.org
Facts on NorduGrid

- A large international Grid facility which originates from North Europe but OPEN to the WORLD
  - Occasionally Japan, Canada, Switzerland appear on the Grid
- An “academic” Grid, serves researchers and consists of academic resources
- A general purpose Grid, it is not a High Energy Physics Grid although at the moment they form the main user group
- A Grid built from “bottom to top”, at the beginning connecting already existing resources belonging to different administrative domains
- Newly acquired dedicated resources are being added
  - SweGrid (600 CPUs), DCGC, etc...
Facts on NorduGrid

- It is a Production Grid and not a “DEMO Grid” nor a Testbed (testbeds have limited lifetime, they are centrally managed and have no real users)
  - 24/7 operation, maintenance
  - real users with real tasks since summer 2002
- An evolving Grid, new middleware versions, new grid services are deployed continuously
- A dynamic Grid, resources come and go
- Mostly a cluster Grid but there'll be some SMPs as well
  - Non-dedicated clusters
  - Homogeneity is preferred within a (sub)cluster
  - Many flavours of Linuxes + some UNIXes
Facts on NorduGrid

- NorduGrid resources cover almost everything the Nordic academics can provide from small to large
  - Small dedicated test clusters (1-4 CPUs)
  - Some junkyard-class second-hand clusters (4 to 80 CPUs)
  - Few university production-class facilities (20 to 60 CPUs)
  - Two world-class TOP500 clusters
    - NSC: Monolith 398 CPUs, 61st position on July/2003 TOP500 list
    - Umeå: Seth 238 CPUs, 203rd position on July/2003 TOP500 list
  - Nordunet Network
  - Storage is added as needed (~10 TB)
Facts on NorduGrid

The Grid runs on a Globus-based middleware
- but it is not a GT2 deployment project, GT2 services are replaced!
  - no Gram!, no Globus-Gatekeeper, no Globus-jobmanager, no Globus MDS schema
- rather, the GT2 libraries and framework were used to create our own middleware with alternative Grid services
  - Grid-manager, Jobplugin, Gridftp server, Userinterface & Broker, Information model, XRSL
- NorduGrid heavily relies on GT2
  - NorduGrid bugreports, critical fixes to Globus
  - Sideeffect: we patch, package and redistribute the GT2 releases (single globus-ng rpm)
  - We've been among the first users and heavy testers since the early GT-2.0 pre-alpha
How does it look like?

Grid Monitor

2003-07-08 CEST 22:58:00

Processes:  Grid  Local

<table>
<thead>
<tr>
<th>Cluster</th>
<th>CPUs</th>
<th>Load (processes: Grid+Local)</th>
<th>Queueing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Denmark</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell (DTU)</td>
<td>32</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td>LSCF</td>
<td>30</td>
<td>18+8</td>
<td>0</td>
</tr>
<tr>
<td>NBI Grid</td>
<td>2</td>
<td>1+8</td>
<td>1</td>
</tr>
<tr>
<td>HEPAX</td>
<td>1</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td><strong>Finland</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hirmu Redhat</td>
<td>2</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td>Hirmu Cluster</td>
<td>10</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td><strong>Japan</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICEPP</td>
<td>1</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td>KEK</td>
<td>1</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td><strong>Norway</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parallab</td>
<td>62</td>
<td>41+11</td>
<td>0</td>
</tr>
<tr>
<td>FI Grid</td>
<td>4</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td>UIO Grid</td>
<td>3</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td><strong>Sweden</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCFAB</td>
<td>19</td>
<td>8+8</td>
<td>46</td>
</tr>
<tr>
<td>Quark Grid</td>
<td>7</td>
<td>6+8</td>
<td>11</td>
</tr>
<tr>
<td>Ingvar (NSC)</td>
<td>31</td>
<td>8+20</td>
<td>3</td>
</tr>
<tr>
<td>Grendel</td>
<td>16</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td>TSL Grid</td>
<td>4</td>
<td>8+8</td>
<td>0</td>
</tr>
<tr>
<td>Monolith (NSC)</td>
<td>396</td>
<td>8+386</td>
<td>353</td>
</tr>
<tr>
<td>Seth (HPC2N)</td>
<td>232</td>
<td>8+211</td>
<td>15</td>
</tr>
<tr>
<td>Old Kosufy</td>
<td>76</td>
<td>8+5</td>
<td>0</td>
</tr>
</tbody>
</table>

**TOTAL**

|         | 19 clusters | 929 | 76 + 633 |

balazs.konya@hep.lu.se
www.nordugrid.org
Understanding NorduGrid: our philosophy

- Create a *functional system* within finite amount of time
- Start with something simple that *works for users* and add functionality gradually
- **Follow standards**, try to reuse available “de facto standard” implementations
- Don't be afraid to **implement your own solution** in case of no standards and if the “de facto standard” software lacks crucial functionalities.
  - GRAM 1.5 and Globus MDS schema is not used within NorduGrid
- Try to be **user friendly**: application driven development
- Try to be **sysadmin friendly**:
  - Don't require full reconfiguration
  - Create light-weight, portable and non-invasive software
  - Support those systems which are needed
Under the hood: the components

NorduGrid Architectural Components

- Control connection
- Data connection

User Interface
Broker

Virtual Organisation database

Globus Replica Catalog

NorduGrid GridManager

XRSL

GSI everywhere

Information System:
network of GIIS/GRIS

Cluster

Worker nodes
Under the hood: How does it work?

- **Information system**, the nervous system of the Grid knows everything
  - Carefully configured stripped down and patched Globus MDS
  - Information is generated on request (efficient providers)
  - Cacheing (only on the resource, GRIS cache)
  - Distributed and multi-rooted: actually a set of linked dynamic LDAP-trees

- The heart(s) of the Grid, the “Grid Manager” on each front-end do most of the job
  - Pre- and post- stages files:
  - Interacts with LRMS, keeps track of job status
  - Manages session directories of grid jobs, cleans up the mess
  - Sends mails to users

- The brain(s) of the Grid, the client “UserInterface” do resource discovery, brokering, Grid job submission and interacts with jobs
  - Interprets user’s job task (XRSL) and performs resource discovery
  - Gets the testbed and job status from the information system
  - “Uploads the job” to the selected resource together with some optional file uploading
  - Job monitoring, termination, retrieval, cleaning, etc
Information System

- Uses Globus’ MDS 2.2
  - Soft-state registration allows creation of any dynamic structure
  - Multi-rooted tree
  - GIIS caching is not used by the clients
  - Several patches and bug fixes are applied
- Information model (schema) represents
  - Clusters
  - Grid jobs
  - Grid users
- Efficient providers and GRIS caching creates a dynamic distributed database
Grid layer on the Front-end: Grid Manager

- Grid Manager replaces Globus’ GRAM, still using Globus Toolkit™ 2 libraries
- All transfers are made via GridFTP
- Added a possibility to pre- and post-stage files, optionally using Replica Catalog information
- Caching of pre-staged files is enabled
- Runtime environment support
- Session directory management
The User Interface

Provides a set of utilities to be invoked from the command line:

- **ngsub** to submit a task
- **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 cancel a job request
- **ngclean** to delete a job from a remote cluster
- **ngrenew** to renew user’s proxy
- **ngsync** to synchronize the local job info with the MDS
- **ngcopy** to transfer files to, from and between clusters
- **ngremove** to remove files

Contains a **broker** that polls MDS and decides to which queue at which cluster a job should be submitted:

- The user must be authorized to use the cluster and the queue
- The cluster’s and queue’s characteristics must match the requirements specified in the xRSL string (max CPU time, required free disk space, installed software etc)
- If the job requires a file that is registered in a Replica Catalog, the brokering gives priority to clusters where a copy of the file is already present
- From all queues that fulfills the criteria one is chosen randomly, with a weight proportional to the number of free CPUs available for the user in each queue
- If there are no available CPUs in any of the queues, the job is submitted to the queue with the lowest number of queued job per processor

Not a centralized service, every user has his/her own personal broker in the UI!
Job Description: extended

Globus RSL

```plaintext
(&{executable="recon.gen.v5.NG"
(arguments="dc1.002000.lumi02.01101.hlt.pythia_jet_17.zebra" "dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.ege7.602.ntuple"
"eg7.602.job" "999"
.stdout="dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.ege7.602.log"
.stdlog="gridlog.txt"(join="yes")

|&{cluster="farm.hep.lu.se"}(cluster="lsf.nbi.dk"(*cluster="seth.hpc2n.umu.se"))(cluster="login-3.monolith.nsc.liu.se")
(inputfiles="dc1.002000.lumi02.01101.hlt.pythia_jet_17.zebra"
"rc://grid.uio.no/lc=dc1.lumi02.002000,rc=NorduGrid,dc=nordugrid,dc=org/zebra/dc1.002000.lumi02.01101.hlt.pythia_jet_17.zebra"
("recon.gen.v5.NG" "http://www.nordugrid.org/applications/dc1/recon/recon.gen.v5.NG.db")
("eg7.602.job" "http://www.nordugrid.org/applications/dc1/recon/eg7.602.job.db")
("noisedb.tgz" "http://www.nordugrid.org/applications/dc1/recon/noisedb.tgz")
)
)
(inputfiles="dc1.002000.lumi02.01101.hlt.pythia_jet_17.zebra"
"rc://grid.uio.no/lc=dc1.lumi02.002000,rc=NorduGrid,dc=nordugrid,dc=org/zebra/dc1.002000.lumi02.01101.hlt.pythia_jet_17.zebra"
("recon.gen.v5.NG" "http://www.nordugrid.org/applications/dc1/recon/recon.gen.v5.NG")
("eg7.602.job" "http://www.nordugrid.org/applications/dc1/recon/eg7.602.job")
)
)
(outputFiles="dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.ege7.602.log"
"rc://grid.uio.no/lc=dc1.lumi02.recon.002000,rc=NorduGrid,dc=nordugrid,dc=org/log/dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.
eg7.602.log")
(histo.hbook"
"rc://grid.uio.no/lc=dc1.lumi02.recon.002000,rc=NorduGrid,dc=nordugrid,dc=org/histo/dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_1
7.ege7.602.histo"
("dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.ege7.602.ntuple"
"rc://grid.uio.no/lc=dc1.lumi02.recon.002000,rc=NorduGrid,dc=nordugrid,dc=org/ntuple/dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_1
7.ege7.602.ntuple")
)
jobname="dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.ege7.602"
(runTimeEnvironment="ATLAS-6.0.2")
(CpuTime=1440)(Disk=3000)(ftpThreads=10))
```
Behind the wheel: Users & Applications

- “We are also grateful to the NorduGrid project, for inviting us test their facilities. Many of our results were obtained with grid computing” T.Sjöstrand and P.Z.Skands, Baryon Number Violation and String Topologies, *Nuclear Physics B* vol. 659, no. 1-2, (2003) pp.243-298


- 76 registered users in the “NorduGrid VO” plus other “alien Vos” and local user lists

- 38 person run at least one job on the Grid in the last month

- Core users: the ATLAS HEP Group and 5-10 heavy users
Behind the wheel: Users & Applications

- **New user areas** which will be supported by the NGDF project:
  - biology, quantum chemistry

- **SweGrid applications** to start in next year (some of the projects which requested CPU time on SweGrid provided resources):
  - “A Novel computational approach to fold recognition”
  - “Numerical simulation of three-dimensional plasma dynamics..”
  - “Modelling the climate of the Baltic Sea”
  - “Quantum Monte Carlo simulations of spin systems”

Please note, that we did not have “dissemination and marketing” team, no manpower to recruit, support, introduce potential new users. Recently started Nordic Grid projects (NGDF, DCGC, SweGrid) are expected to fill this gap.
Behind the wheel: HEP Atlas group

- The Scandinavian High Energy Physics Atlas Group has been the driving force behind the NorduGrid development
- Atlas DATA Challenges
  - A series of computing challenges of the Atlas collaboration with increasing size and complexity
  - Preparation for the 2007 start of the Large Hadron Collider
  - Around 50 institutes from all around the world
  - Scandinavia participated via the NorduGrid
  - The NorduGrid contribution ramped up from 2% to 15% within a year due to expansion (ca. 4TB of data, more than 4750 single jobs 2 to 30 hours long)
Behind the wheel: what does a user get?

- A solid distributed infrastructure
  - Uniform Grid layer on top of clusters
  - Uniform security infrastructure
  - Grid-aware storage
  - Supported Runtime environments

- As soon as political issues are solved users “enjoy the power” of a Production Grid:
  - Users somehow have to be related to the resources
  - No free lunch (resources) any longer

- A user friendly interface (requires some learning)
  - Powerful command line User Interface
  - Or a Gridportal: GridBlocks
Missing functionalities

- Authorization, Authorization, Authorization
  - User management (groups): VOMS
  - Access to data: GACL
  - Information System

- The infrastructure of a non-free (commercial) Grid
  - Policies, Accounting, Grid Market

- Failure management

- Scalable and reliable indexing services for
  - Information system, data management

- Any new development will be in the Grid Services framework

- These are nontrivial challenges, most of them are not just standardization problems!
Summary

- There is Production Grid in North Europe since summer 2002
- The Grid runs on an alternative middleware solution which was developed by making use of the Globus Toolkit 2 libraries
- The middleware reliably implements the core Grid services such as job submission, job management, resource discovery, brokering, basic data management
- Still a lot needs to be done, especially in the Policies, Authorization area
- The middleware is publicly available, the development follows an open source track
- The Grid is OPEN for participation
Further info, contacts

- **www.nordugrid.org**
  - Grid Monitor: in case you are curious if anybody is using the Grid ;)
  - Downloads: middleware source & binaries, setup instructions plus the globus-nordugrid-rpm ;)
  - Documentation: papers, presentations, manuals and even tutorial material
  - Contacts: if you are looking for cooperation (you may check out the tasklist as well)
  - Mailing lists: nordugrid-discuss & nordugrid-support

- **Have a Testdrive on the NorduGrid!**
  - I am available during SC2003 and ready to give a live demonstration at any time (just drop an email).
acknowledgments

- Developers: A.Konstantinov, M.Ellert, O. Smirnova, J. L. Nielsen, A. Wäänänen, B.Kónya, T.Myklebust, J.Herrala, M.Tuisku
- System administrators from the centers: Ulf Mjörnmark and Björn Lundberg of Lund University, Björn Nilsson of NBI Copenhagen, Niclas Andersson and Leif Nixon of NSC Linköping, Åke Sandgren of HPC2N Umeå and Jacko Koster from Paralab
- Some of our users: Alex Read, P.Z.Skands, Nils Gollub, O.F.Syljuasen
- Management, Steering Committee: P. Eerola, T.Ekelöf, J.R.Hansen, F.Ould-Saada, B.Vinter