Next Generation of Grid Services for the NorduGrid

NEC 2003
Varna, September 19, 2003

Oxana Smirnova
Outline

- Project overview
- Services in details
- Present vs future
NorduGrid: some facts

- **NorduGrid is:**
  - A Globus-based Grid middleware solution of choice in Scandinavia and Finland
  - A large international 24/7 production quality Grid facility
  - A resource routinely used by researchers since summer 2002
  - A freely available software
  - A project in development

- **NorduGrid is NOT:**
  - Based on EDG/LCG solutions
  - ATLAS- or HEP-dedicated tool
  - A *testbed* anymore
  - A finalized solution

2003-09-19

oxana.smirnova@hep.lu.se
www.nordugrid.org
What makes NorduGrid different from other Grids

1. It is stable by design:
   a) The nervous system: distributed yet stable Information System (Globus’ MDS 2.2+patches)
   b) The heart(s): Grid Manager, the service to be installed at master nodes (based on Globus, replaces GRAM)
   c) The brain(s): User Interface, the client/broker that can be installed anywhere as a standalone module (makes use of Globus)

2. It is light-weight, portable and non-invasive:
   a) Resource owners retain full control; Grid Manager is effectively a yet another user (with many faces though)
   b) No requirements w.r.t. OS, resource configuration, etc.
   c) Clusters need not be dedicated
   d) Runs on top of existing Globus installation (e.g. VDT)

3. Strategy: start with something simple that works for users and add functionality gradually
Current standing

- Project involves several **Nordic universities** and HPC centers
- Will continue for 4-5 years more
  - Forms the "**North European Grid Federation**" of the **EGEE** together with the Dutch Grid, Belgium and Estonia
  - Will provide middleware for the "**Nordic Data Grid Facility**" and related projects (SWEGRID, Danish Grid etc)
- Shares authentication and authorization mechanisms with EDG/LCG-1/EGEE
- The **first milestone** reached in time: it provides essential facilities for the ATLAS Data Challenge (up to 15% contribution during the summer holidays)
- Current stable release: 0.3.26
- **Next milestone**: the first release, NorduGrid 1.0
- **Any new development will be in the Grid Services framework**
The resources

- Almost everything the Nordic academics can provide (ca 1000 CPUs in total):
  - 4 dedicated test clusters (3-4 CPUs)
  - Some junkyard-class second-hand clusters (4 to 80 CPUs)
  - Few university production-class facilities (20 to 60 CPUs)
  - Two world-class clusters in Sweden, listed in Top500 (238 and 398 CPUs)

- Other resources come and go
  - Canada, Japan – test set-ups
  - CERN, Dubna – clients
  - It’s open, anybody can join or part

- People:
  - the “core” team: 6 persons
  - local sysadmins are only called up when users need an upgrade
A snapshot

Grid Monitor

2003-07-08 CEST 22:58:00

<table>
<thead>
<tr>
<th>Cluster</th>
<th>CPUs</th>
<th>Load (processes: Grid-local)</th>
<th>Queueing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell (DTU)</td>
<td>32</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>LSCF</td>
<td>30</td>
<td>18.0</td>
<td>0</td>
</tr>
<tr>
<td>NBI Grid</td>
<td>2</td>
<td>1.0</td>
<td>1</td>
</tr>
<tr>
<td>HEPAX</td>
<td>1</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Finland: Hirmu Redhat</td>
<td>2</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Finland: Hirmu Cluster</td>
<td>10</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Japan: ICEPP</td>
<td>1</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Japan: KLK</td>
<td>1</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Norway: Parallab</td>
<td>62</td>
<td>41.0</td>
<td>0</td>
</tr>
<tr>
<td>Norway: FT Grid</td>
<td>4</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Norway: UIO Grid</td>
<td>3</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td>Sweden: SCFAB</td>
<td>19</td>
<td>0.0</td>
<td>46</td>
</tr>
<tr>
<td>Sweden: Quark Grid</td>
<td>7</td>
<td>5.0</td>
<td>11</td>
</tr>
<tr>
<td>Sweden: Ingvar (NSC)</td>
<td>31</td>
<td>0.0</td>
<td>3</td>
</tr>
<tr>
<td>Sweden: Grendel</td>
<td>16</td>
<td>2.0</td>
<td>0</td>
</tr>
<tr>
<td>Sweden: TSL Grid</td>
<td>4</td>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>Sweden: Monolith (NSC)</td>
<td>395</td>
<td>0.0</td>
<td>353</td>
</tr>
<tr>
<td>Sweden: Seth (HPC2N)</td>
<td>232</td>
<td>0.0</td>
<td>15</td>
</tr>
<tr>
<td>Sweden: Old Kosufy</td>
<td>76</td>
<td>0.0</td>
<td>0</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>19</td>
<td>929</td>
<td>76 + 633</td>
</tr>
</tbody>
</table>

2003-09-19

oxana.smirnova@hep.lu.se

www.nordugrid.org
Components
How does it work

- **Information system** knows everything
  - Substantially re-worked and patched Globus MDS
  - Distributed and multi-rooted
  - Allows for a mesh topology
  - No need for a centralized broker

- The server ("*Grid manager*”) on each gatekeeper does most of the job
  - Pre- and post- stages files
  - Interacts with LRMS
  - Keeps track of job status
  - Cleans up the mess
  - Sends mails to users

- The client ("*User Interface*”) does the brokering, Grid job submission, monitoring, termination, retrieval, cleaning etc
  - Interprets user’s job task
  - Gets the testbed status from the information system
  - Forwards the task to the best Grid Manager
  - Does some file uploading, if requested
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

- A new schema is developed, to serve clusters
  - Clusters are expected to be fairly homogeneous
Front-end and the 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

![Diagram of Grid System](image_url)
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
Job Description: extended
Globus RSL

(&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.eg7.602.ntuple"
 "eg7.602.job" "999")
(stdin="dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.eg7.602.log")
(stdlog="gridlog.txt")[join="yes")

)(&((cluster="farm.hep.lu.se")((cluster="lscf.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"))

(outputFiles= ("dc1.002000.lumi02.01101.hlt.pythia_jet_17.eg7.602.log"
 "rc://grid.uio.no/lc=dc1.lumi02.recon.007.01101,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.007.01101,rc=NorduGrid,dc=nordugrid,dc=org/histo/dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.eg7.602.histo")
 ("dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.eg7.602.ntuple"
 "rc://grid.uio.no/lc=dc1.lumi02.recon.007.01101,rc=NorduGrid,dc=nordugrid,dc=org/ntuple/dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.eg7.602.ntuple")
 (jobname="dc1.002000.lumi02.recon.007.01101.hlt.pythia_jet_17.eg7.602")
 (runTimeEnvironment="ATLAS-6.0.2")
 (CpuTime=1440)(Disk=3000)(ftpThreads=10))
Task flow

Grid

Manager

Gatekeeper

GridFTP

Front-end

Cluster

RC

SE

UI+RB

MDS

RSL

oxana.smirnova@hep.lu.se

www.nordugrid.org
What NorduGrid 1.0 will provide…
1. Stability, reliability

- The facility works non-stop since May 2001
- Grid components do not require special attention of sysadmins
- Failure of a single cluster/node does not cripple the whole system
- Comes complete with a test-suite
2. Portability

- Runs on any Linux flavor and Solaris
  - Tests run on Tru64
  - Sources are freely available to port anywhere
- Does not require very special cluster configuration
  - Nothing needs to be installed on worker nodes, only on a gatekeeper
3. Smart front-end and client

- **Grid Manager** replaces Globus GRAM
  - All transfers are done via GridFTP
  - Pre- and post-staging of files (even using other protocols)
- Caching of pre-staged files
- Support for runtime environment
- Extended RSL
- Decentralized resource brokering by each client
  - Relies on Information system
4. User-oriented services

- Support for a variety of applications
  - ATLAS is the major customer
    - Data Challenges: of 19 ATLAS “sites”, ramped up contribution from 2% to 15% within a year due to expansion (ca. 4TB of data, more than 4750 single jobs 2 to 30 hours long)
    - Higgs and B-physics team: own Data Challenges
  - Combined LEP analysis
  - Theoretical high energy physics: tests of PYTHIA components
  - Quantum lattice models

- Simple monitoring
  - Users can make own choice of resources
NorduGrid 1.0 issues (to be addressed by next releases):
1. Scalability

- Only 8 production-level clusters have the NorduGrid installed
  - There are no dedicated clusters available, hence Grid users have to compete for the resources
  - Only half of the owners let all the resources to be used by the Grid
- Only 2-3 regular users (ATLAS) and ~5-7 occasional ones
  - That’s still a lot given the small size of the Scandinavian physics community
- Maximum observed load is only about 160 concurrent jobs, plus ca. 200 queued
  - This is due to the internal limits and priorities set by resource owners
2. QoS assurance

- Clusters can provide false information and yet stay in the system
  - The issue is solvable technically, but not politically
- Clusters can provide poor service, such as not updating credentials or experiencing local hardware problems causing massive job failures
  - and yet they can not be suspended/banned
  - Also a political issue, not technical
- No warnings prior to service and users credentials expiration
- No job recovery and/or rescheduling mechanisms
3. Logging, accounting, auditing

- Transient nature of all the information
  - Very difficult to trace back eventual problems
  - Impossible to provide precise usage figures
  - Very difficult to keep track of who done what

- General lack of corresponding services
  - Unfortunately, this is an issue for any Grid solution worldwide, maybe EGEE could help?
  - Meanwhile, a “logger” service is being slowly introduced
4. Authorization and authentication

- As every Grid project worldwide, NorduGrid uses rudimentary solutions from Globus2:
  - No Virtual Organization management
  - No VO-based authorization or access rights control
  - Insufficient and ambiguous information stored in certificates
  - No automatic certificate/key download by jobs
  - No proxy management
    - No secure proxy storing/delegating service
    - No automatic proxy extension/renewal

- These are also the issues which should be addressed globally (GGF, EGEE?)
  - VOMS is being implemented by NorduGrid, EDG/LCG-1
5. Information system

- Globus MDS is known to be unreliable
  - NorduGrid has provided several patches, but the implementation is still unsatisfactory

- Has hierarchical rather than mesh structure
  - A single failed machine can deem entire country resources unavailable

- Does not use caching
  - Frequent timeouts

- Anonymous access
  - Potentially sensitive information is unprotected
6. Data management system

- There is no DMS satisfying a HEP experiment requirements on the market – a global problem
- NorduGrid supports only most basic Globus Replica Catalog functionality
  - No Storage Element concept
  - No mass storage support
  - Centralized – single point of failure
  - Slow
  - Questionable scalability
  - No VO-based access control
- A “Clever SE” is being developed
7. Front-end functionality

- Used GridFTP protocol lacks a lot of functionality
- So far, is interfaced only to PBS
- A shared file system between the front-end and the worker nodes is still required
- No interactivity
  - Does not provide any channel between a running job and a user
8. Workload management

- Resource brokering is nothing more than user request matching
  - No pre-allocation
  - No cost evaluation
  - No load balancing
- No rescheduling and resubmission
  - Should rely on the “logger”
- No VO-based management
- No cross-cluster parallelism
9. Installation and configuration

- Provided in a half a dozen of separate packages (tarballs or RPMs)
  - Plus twice that amount of external software like patched Globus, patched perl-ldap etc.
- No installer provided
- No installation instructions provided
- Configuration is a very time-consuming task
  - XML-based configuration procedure is being considered
10. Manpower

- No full-time developers at the moment
  - Bug fixes is the major activity before the first release
  - 130+ bugs and enhancement requests registered since May 2003
- No researchers from computer sciences are volunteering to get involved
- Reduced user support
- Some contribution via EGEE is anticipated
Summary

- NorduGrid pre-release (currently 0.3.26) **works reliably**
- Release 1.0 is slowly but surely on its way; many fixes are needed
- Much functionality will still be missing from 1.0
  - Most of that functionality is missing from other Grids as well
  - the problem is global
- We welcome new users and resources
  - As soon as political issues are solved, users “enjoy the power”
  - Beware, support is somewhat limited at the moment
- Future: Nordic countries are committed to use and support NorduGrid
  - Some resources (Sweden) have to use LCG-1, but plan to use NorduGrid on top (or in parallel)
  - The project is open for members and contributions from everywhere