Distributed training and testing
grid infrastructure evolution

Nikolay.Kutovskiy@jinr.ru

Laboratory of information technologies, Joint institute for nuclear research
(Dubna, Russia)
National Scientific and Educational Centre of Particle and High Energy
Physics of the Belarusian State University (Minsk, Belarus)

GRID'2012, 16 - 21.07.2012 (JINR, Dubna, Russia)
Motivation

Grid in science

HEP
biology
medicine
materials

Grid in business

industry
finance
media
logistics

money
anti-laundering
Motivation

- Grid has already become a standard tool used in different fields of human activity.
- There is a growing demand in such specialists who could use grid, deploy and maintain grid sites and services as well as develop new components and port applications to run in grid environments.
- Production grid infrastructures can't be used for trainings as well as for middleware development, testing and evaluation because of the following reasons:
  - fast response time is required (e.g. to get a result of submitted jobs asap),
  - provide superuser privileges for sysadmin trainings,
  - etc.
That's why a dedicated distributed training and testing grid-infrastructure (t-infrastructure for short) had been deployed. Since performance is not a critical issue for tasks it is aimed for all its services are run inside VMs. t-infrastructure is used for giving trainings for students of University Centre of JINR, University «Dubna», colleagues from JINR and its member states as well as for performing obligations in different Grid related activities of local and international projects.
JINR components of t-infrastructure

- 3 gLite/EMI-based grid sites;
- GT5-based testbed;
- BOINC server, application repository, EDGI-bridge;
- WLCG testbed;
- tier3mon testbeds;
- development web-GUI for POI of RGN;
- server with gLite API for grid services development;
- web-server for web-portal.
gLite/EMI-based testbed

- **Trainings:**
  - users,
  - system administrators,
  - developers.

- **WLCG related activities:**
  - evaluation, bugs submission and patches certification for MPI enabled CEs.
  - functional tests for LFC perl API.
gLite/EMI-based part of t-infrastructure (schema)

Added in 2011

2012?

BY-GSTU

AZ-IP

Added in 2011
<table>
<thead>
<tr>
<th>site name</th>
<th>hosting organization</th>
<th>services</th>
</tr>
</thead>
<tbody>
<tr>
<td>RU-JINR</td>
<td>JINR (Dubna, Russia)</td>
<td>UI, LCG-CE + 2 WNs, DPM SE, LFC, WMS, LB,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sBDII, tBDII, VOMS</td>
</tr>
<tr>
<td>RU-JINR-2</td>
<td>– / –</td>
<td>LCG-CE + 2 WNs, DPM SE, sBDII</td>
</tr>
<tr>
<td>RU-JINR-MPI</td>
<td>– / –</td>
<td>MPI-enabled CREAM + 3 WNs, DPM SE, sBDII</td>
</tr>
<tr>
<td>BG-SU</td>
<td>Sofia University “St. Kliment Ohridski”</td>
<td>UI, LCG-CE + 4 WNs, DPM SE, sBDII</td>
</tr>
<tr>
<td></td>
<td>(Sofia, Bulgaria)</td>
<td></td>
</tr>
<tr>
<td>SU-Protvino-IHEP</td>
<td>Institute of High-Energy Physics (Protvino,</td>
<td>UI, LCG-CE + 2 WNs, dCache SE, WMS, LB, sBDII</td>
</tr>
<tr>
<td></td>
<td>Russia)</td>
<td></td>
</tr>
<tr>
<td>UZ-IMIT</td>
<td>Institute of Mathematics and Information</td>
<td>UI, MPI enabled LCG-CE + 4 WNs, WMS, LB, sBDII</td>
</tr>
<tr>
<td></td>
<td>technologies (Tashkhtent, Uzbekistan)</td>
<td></td>
</tr>
<tr>
<td>UA-BITP</td>
<td>Bogolyubov Institute for Theoretical Physics</td>
<td>UI, LCG-CE + 8 WNs, DPM SE, LFC, WMS, LB, sBDII</td>
</tr>
<tr>
<td></td>
<td>(Kiev, Ukraine)</td>
<td></td>
</tr>
<tr>
<td>UA-KPI-HPCC</td>
<td>National Technical University &quot;Kyiv UI,</td>
<td>UI, LCG-CE + 8 WNs, DPM SE, LFC, WMS, LB, sBDII</td>
</tr>
<tr>
<td></td>
<td>LCG-CE + 8 WNs, DPM SE, LFC, WMS, LB, sBDII</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polytechnic Institute&quot; (Kiev, Ukraine)</td>
<td></td>
</tr>
<tr>
<td>KZ-ENU</td>
<td>L.N. Gumilyov Eurasian National University</td>
<td>UI, CREAM + 4 WNs, sBDII, DPM SE</td>
</tr>
<tr>
<td></td>
<td>(Astana, Kazakhstan)</td>
<td></td>
</tr>
<tr>
<td>UA-ILTPE</td>
<td>B.Verkin Institute for Low Temperature Physics</td>
<td>UI, CREAM + 2 WNs, sBDII</td>
</tr>
<tr>
<td></td>
<td>and Engineering (Kharkov, Ukraine)</td>
<td></td>
</tr>
</tbody>
</table>
Training courses, tutorials and lectures for users

- Semestral educational course at the University “Dubna”:
  - 5 years of successful experience, more than 350 students;
  - Grid basics;
  - gLite middleware;
  - practical work within the Grid;
  - scientific research activities of students and post-graduate students.

- Semestral educational course at the JINR University Centre:
  - focus on research activities;
  - train physicists for their practical work within the Grid.

- Introduction lectures and short-term courses for participants of different international schools and practices:
  - “JINR /CERN grid and advanced information systems” schools,
  - practice for students from Arab Republic of Egypt,
  - introduction lecture for participants of training of young scientists from CIS, Russian student scientific and technical school “Personnel of the future” and more.
Trainings for system administrators

In person:

1) at JINR
2) at host organisation

Remotely:

audio/video-conferences
text (e.g. support via email, chat)
Trainings for system administrators

- Each trainee must have a dedicated resources with super-user privileges.
- JINR resources or trainee's home organisation resources can be used:
  - JINR resources:
    - bunch of VMs for grid services deployment,
    - life-time of such resources = training duration.
  - home organisation resources:
    - VMs or physical servers,
    - grid site deployed during training can be used later on.
- Trainings also vary in duration:
  - short-term (intensive: a several full days in a row guided by trainer, in person or in video-conference mode),
  - medium-term (a few weeks – month: half of the day – guided by trainer in person or in video-conference mode, the rest of the day - self-guided work),
  - long-term (months: support via email/chat).
Conducted sysadmins trainings/support

<table>
<thead>
<tr>
<th>Organisation</th>
<th>City</th>
<th>Country</th>
<th>Dates</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>BITP, KPI</td>
<td>Kiev</td>
<td>Ukraine</td>
<td>05.09.11 - 09.09.11</td>
<td>in person</td>
</tr>
<tr>
<td>IP</td>
<td>Baku</td>
<td>Azerbaijan</td>
<td>13.07.11 - 31.08.11</td>
<td>in person</td>
</tr>
<tr>
<td>ENU</td>
<td>Astana</td>
<td>Kazakhstan</td>
<td>14.06.11 - 10.07.11</td>
<td>in person</td>
</tr>
<tr>
<td>NUM, SICT of MUST</td>
<td>Ulaanbaatar</td>
<td>Mongolia</td>
<td>17.05.11 - 20.06.11</td>
<td>in person</td>
</tr>
<tr>
<td>ILTPE</td>
<td>Kharkov</td>
<td>Ukraine</td>
<td>06.12.10 - ...</td>
<td>email</td>
</tr>
<tr>
<td>IMIT</td>
<td>Tashkent</td>
<td>Uzbekistan</td>
<td>30.11.09 - 13.12.09</td>
<td>in person</td>
</tr>
<tr>
<td>BITP, KPI</td>
<td>Kiev</td>
<td>Ukraine</td>
<td>09.11.09 - 27.11.09</td>
<td>video-conf</td>
</tr>
<tr>
<td>NIRDIMT</td>
<td>Cluj</td>
<td>Romania</td>
<td>15.10.09 - 30.10.09</td>
<td>in person</td>
</tr>
<tr>
<td>UCT</td>
<td>Cape Town</td>
<td>South Africa</td>
<td>01.06.08 - 13.06.08</td>
<td>in person</td>
</tr>
<tr>
<td>NC PHEP BSU</td>
<td>Minsk</td>
<td>Belarus</td>
<td>2008</td>
<td>email</td>
</tr>
</tbody>
</table>
grid sites of organisations from JINR member states

- trained administrators deploy grid sites at home organisations what can become a part of one of the global grid infrastructure (e.g. WLCG) if grid site matches a set of requirements such as internet connectivity, computational and storage resources, ability to provide a required rate of reliability and availability.
- if the grid site doesn't match at least one of the mentioned requirements and hence can't be a part of global grid infrastructure then it can be integrated into t-infrastructure what doesn't have any of such demands.
- As soon as grid site starts matching all requirements it can be reconfigured to become a part of the global grid infrastructure.
grid sites of organisations from JINR member states

production grid infrastructures

matches

grid site

requirements in internet connectivity, computational and storage resources, reliability, availability

doesn't match yet

t-infrastructure
There are no regular trainings for developers of grid services and grid-enabled applications yet.

They are trained upon necessity for particular project.
gLite/EMI-based testbed

- Traings:
  - users,
  - system administrators,
  - developers.

- **WLCG related activities:**
  - evaluation, bugs submittion and patches certification for MPI enabled CEs.
  - functional tests for LFC perl API.
Monitoring tools development for ATLAS tier 3 sites (1/5)

- Several types of local resources management systems (LRMS) and mass storage systems (MSS) are used on ATLAS tier3 sites:
  - LRMS: PROOF, PBS, Condor, Oracle Grid Engine, LSF.
  - MSS: XRootD, dCache, DPM, NFS, GPFS, Lustre.
- A development of software suite for local site monitoring assumes the following activities:
  - validation of existing monitoring tools for each of the using component,
  - development and debugging the new monitoring tools.
- Activities listed above implies the following:
  - deployment of a separate testbed for each of the LRMS and MSS reported as being used at ATLAS Tier3 sites,
  - ganglia server deployment,
  - ganglia agents installation and configuration for a specific testbed,
  - installation and validation of the additional ganglia plug-ins for monitoring metrics collection.
Monitoring tools development for ATLAS tier 3 sites (2/5)

Development of a software suite for the local site monitoring

validation of the existing monitoring tools for each of the component in use

development and debugging new monitoring tools

deployment of a separate testbed for each of the LRMS and MSS reported as being used at ATLAS Tier-3 sites

Ganglia server deployment

Ganglia agents installation and configuration for a specific testbed

installation and validation of the additional plug-ins for monitoring metrics collection
Monitoring tools development for ATLAS tier 3 sites (3/5)

24/7 availability of the testbeds components with different LRMS and MSS running in parallel on dedicated physical servers would cause a sufficient hardware capacity.

Monitoring tools deployment and development as well as testbeds operation may require redeployment of a certain testbed or its parts.

Testbeds performance is not a critical issue for such tasks.

Virtualization:
- More effective utilization of the hardware resources,
- Ability to perform quickly and easily such operations as VMs creation from existing images/templates, VMs backup before significant changes and VMs restoration from backup if needed.

- All components of each testbed can be run on Linux (inside VM) as well as a physical server.
- Most components do not require own kernel extensions and thus the OS-level virtualization can be used which is more lightweight and faster than full hardware emulation or paravirtualization approaches.
- But there are still some components which require own kernel extensions (e.g. Lustre, GPFS).
Monitoring tools development for ATLAS tier 3 sites (4/5)

- 2 worker nodes
  - Front-end + worker node (3 WNs in total)
  - PBS (torque)

- 2 worker nodes
  - Front-end + worker node (3 WNs in total)
  - PROOF

- 2 worker nodes
  - Front-end + worker node (3 WNs in total)
  - Condor

- 2 worker nodes
  - Front-end (2 WNs in total)
  - OGE

- 2 servers
  - Manager (redirector)
  - XRootD testbed 1

- 2 servers
  - Manager (redirector)
  - XRootD testbed 2

- OpenVZ-enabled server

- 2 servers
  - Ganglia
  - web frontend
  - Development host

- Lustre

- MDS

- OSS

- Xen-enabled server

client
## Monitoring tools development for ATLAS tier 3 sites (5/5)

<table>
<thead>
<tr>
<th>Testbed</th>
<th>Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBS</td>
<td>torque headnode (HN) + worker node (WN) + ganglia (gmond, gmetad, webfronted) + jobmonarch, 2 torque WNs + gmond</td>
</tr>
<tr>
<td>PROOF</td>
<td>HN + gmond, 2 WNs + gmond</td>
</tr>
<tr>
<td>Condor</td>
<td>HN + WN + gmond, WN + gmond, client + gmond</td>
</tr>
<tr>
<td>OGE</td>
<td>HN + Ganglia (gmond, gmetad, webfrontend), 2 WNs + gmond</td>
</tr>
<tr>
<td>XRootD 1</td>
<td>manager + gmond, server + gmond, server + ganglia (gmond, gmetad, webfrontend)</td>
</tr>
<tr>
<td>XRootD 2</td>
<td>manager + gmond, server + gmond, server + ganglia (gmond, gmetad, webfrontend)</td>
</tr>
<tr>
<td>Lustre</td>
<td>MDS + gmond + gmetad + gweb + t3mon-site-lustre, OSS + gmond, client + gmond</td>
</tr>
</tbody>
</table>
To simplify user operations to work with particular application in RNGN problem-oriented web based interfaces were developed by LIT JINR team for the following apps:

- DL_POLY (molecular dynamics);
- Elmer (computer-aided engineering);
- GEANT4-DNA (to simulate biological damages induced by ionising radiation at the cellular and sub-cellular scale).

POIs were developed as plug-ins for RNGN graphical web-based interface what was deployed on t-infrastructure and used for development and debugging.
Private IaaS

- t-infrastructure is rapidly developed because of being used for more and more projects.
- Number of VMs varies from tens to hundred => hard to manage all of them manually => migrate to cloud (IaaS).
- OpenNebula was chosen as a platform for private cloud infrastructure.
- Most of the VMs are based on OpenVZ (apart from 3 Xen VMs).
- OpenNebula doesn't support OpenVZ hypervisor.
- OpenVZ driver for OpenNebula v2.2.1 had been developed in 2011 and recently updated by KPI team to the latest stable OpenNebula version 3.6.
To use idle CPUs of desktop PCs at JINR and organisations from its member states for computational tasks one needs to do the following:

- build DG infrastructure,
- adopt applications to it.

Last item requires a testbed which can be deployed on its infrastructure and integrated within its gLite/EMI-based segment.
First attempt to build local DG testbed had been made in 2010.

A small DG testbed was deployed (BOINC server and few BOINC clients on UC PCs.

To intergrate DG and gLite/EMI testbeds a few more services need to be setup:
- bridge (3G-bridge),
- application repository (AR),
- gridftp-server.

Due to incomplete documentation and luck of support from developers there is not much success yet.
Future plans

- Build private cloud and migrate all t-infrastructure services into it to simplify VMs management across servers.
- Finish the integration of DG and gLite/EMI-based segments.
- Upgrade grid services to their latest stable versions.
- Install WS-PGRADE portal to provide a web-GUI for users in addition to console.
- Use in other projects.
The ultimate goal is to have infrastructure what could become a platform for training, research, development, tests and evaluation of modern technologies for distributed computing and data management.
web-portal on t-infrastructure

- URL: https://gridedu.jinr.ru
- t-infrastructure detailed description,
- tasks it is used for,
- news,
- current state,
- different instructions/howtos/guides for sysadmins,
- etc...
Acknowledgement

- All t-infrastructure related work in 2010 - 2012 years had been supported by JINR grant for young scientists and specialists.