VM-based infrastructure for simulating different cluster and storage solutions used on ATLAS Tier-3 sites

S. Belov[1], I. Kadochnikov[1], V. Korenkov[1], M. Kutouski[1,2] , D. Oleynik[1], A. Petrosyan[1] on behalf of the ATLAS Collaboration

[1] Laboratory of Information Technologies, Joint Institute for Nuclear Research, 141980, Dubna, Russia

[2] National Scientific and Educational Centre of Particle and High Energy Physics of the Belarusian State University, 220040, Minsk, Belarus

E-mail: kut@jinr.ru

Abstract. The current ATLAS Tier-3 infrastructure consists of a variety of sites of different sizes and with a mix of local resource management systems (LRMS) and mass storage system (MSS) implementations. The Tier-3 monitoring suite, having been developed in order to satisfy the needs of Tier-3 site administrators and to aggregate Tier-3 monitoring information on the global VO level, needs to be validated for various combinations of LRMS and MSS solutions along with the corresponding Ganglia plugins. For this purpose the testbed infrastructure, which allows simulation of various computational cluster and storage solutions, had been set up at JINR (Dubna, Russia). This infrastructure provides the ability to run testbeds with various LRMS and MSS implementations, and with the capability to quickly redeploy particular testbeds or their components. Performance of specific components is not a critical issue for development and validation, whereas easy management and deployment are crucial. Therefore virtual machines were chosen for implementation of the validation infrastructure which, though initially developed for Tier-3 monitoring project, can be exploited for other purposes. Load generators for simulation of the computing activities at the farm were developed as a part of this task. The paper will cover concrete implementation, including deployment scenarios, hypervisor details and load simulators.

1. Tasks

According to ATLAS Tier-3 sites survey a several types of local resources management systems (LRMS) and mass storage systems (MSS) are used on ATLAS Tier-3 sites:

- LRMS: PROOF [1], PBS [2], Condor [3], Oracle Grid Engine (OGE) [4], LSF [5].
- MSS: XRootD [6], dCache [7], DPM [8], NFS [9], GPFS [10], Lustre [11].

A development of software suite for local site monitoring assumes the following activities:

- validation of the existing monitoring tools for each of the component in use,
- development and debugging new monitoring tools.

Activities listed above imply the following:

- deployment of a separate testbed for each of the LRMS and MSS reported as being used on ATLAS Tier-3 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 as well as non-related to Ganglia monitoring tools.

2. Virtualization

Testbeds with different LRMS and MSS have to be run in parallel and must be available at any time. Running each component of each testbed on dedicated physical server would require using a substantial amount of hardware.

Monitoring tools deployment and development as well as testbeds operation may require re-deployment of a certain testbed or its parts. Hence there should be the ability to easily manage testbed parts. Testbeds performance is not a critical issue for such task as development. Taking all mentioned above into account, one can conclude that testbeds can be run on virtual machines.

Virtualization allows more effective hardware resources utilization and provides the ability to perform the following operations quickly and easily:

- create VM from existing image/template,
- backup VM before significant changes and quickly restore VM from backup if needed.

All components of each testbed can be run on linux (inside VMs) as well as physical servers themselves. Most of the components do not require kernel extensions. 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).

Among of possible candidates the OpenVZ [12] as a solution for virtualization on OS-level and Xen [13] as a product providing full hardware emulation were chosen because of the following reasons:

- stable and actively developing software with sufficient tool set for VMs management and monitoring,
- strong and helpful community,
- good documentation,
- free software (GNU GPL license).

Apart from that the services of the grid infrastructure for training and testing (see [14] for more details) have been successfully running on OpenVZ-based VMs since 2006.

3. Testbed structure

A deployment of the JINR testbed for ATLAS Tier-3 sites monitoring tools development [15] started in February 2011 as a part of the JINR training and testing grid infrastructure and for the time being the following LRMS and MSS are running on it (see table 1 and figure 1): PBS, Condor, XRootD, Lustre, PROOF, OGE. Apart from that the Ganglia and development servers are deployed. All testbeds excluding Lustre are running on single OpenVZ-enabled server. Lustre services are deployed on Xen-based VMs on a separate server.

Table 1. A list of running services of the LRMS and MSS testbeds

testbed name	services
PBS	torque headnode (HN) + worker node (WN) + Ganglia (gmond, gmetad, webfronted) + jobmonarch, 2 torque WNs + gmond
PROOF	HN + gmond, 2 WNs + gmond
Condor	HN + WN + gmond, WN + gmond, client + gmond
OGE	HN + Ganglia (gmond, gmetad, webfrontend), 2 WNs + gmond

testbed name	services
XRootD 1	manager + gmond, server + gmond, server + gmond
XRootD 2	manager + gmond, server + gmond, server + gmond
Lustre	MDS + gmond + gmetad + gweb + t3mon-site-lustre, OSS + gmond, client + gmond

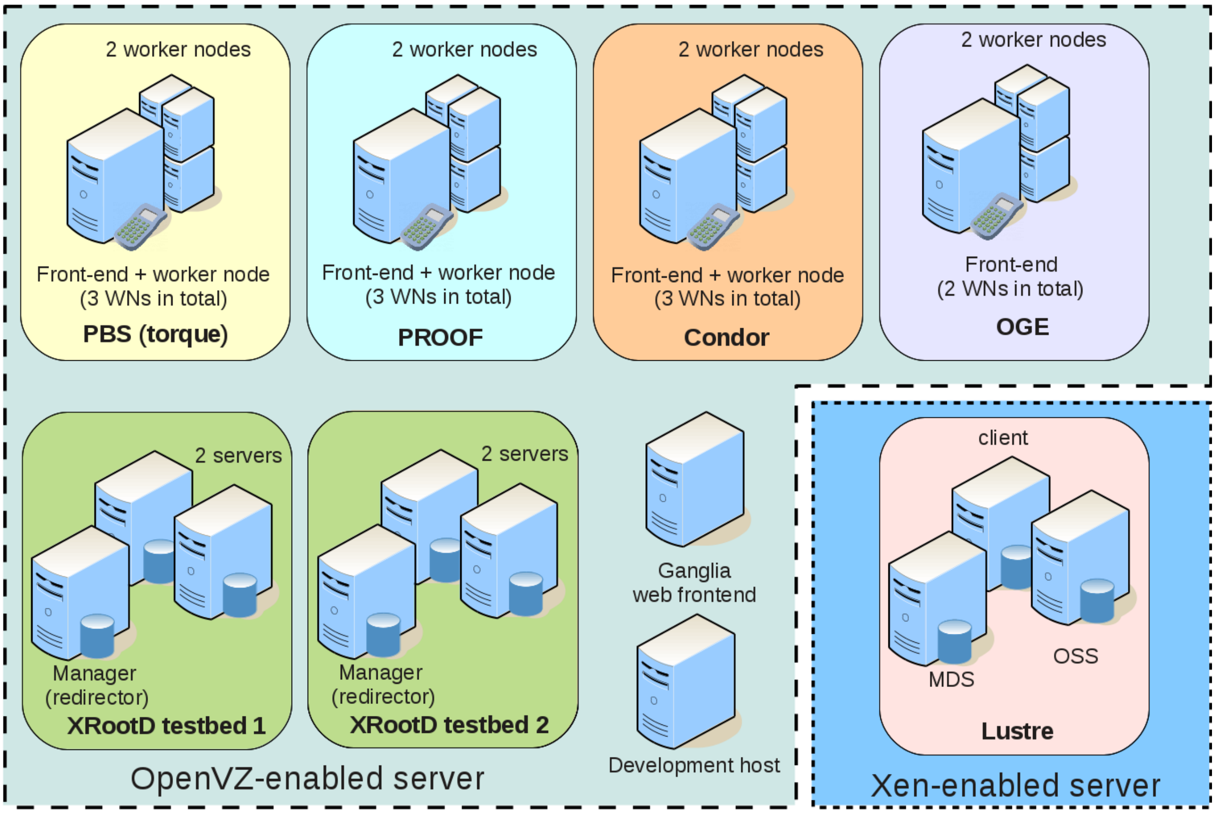

Figure 1. A scheme of the LRMS and MSS testbeds distribution over the servers

4. Load test suite

Such tools as collectors and user interfaces as well as all the monitoring data processing chain should be properly verified and tested. Jobs and files events are the main input data for the monitoring, and on real sites they are normally produced by real users. Speaking of the development testbed, there are a lot of different data collectors and several web interfaces to be continuously tested. Therefore automated production of “user-like” activity events could be very helpful on testbeds.

Kinds of activities which could be used to simulate users’ analysis work:

- job events:
  - random submissions with configurable frequency;
  - adjustable memory usage;
  - CPU load.
- files related events:
  - uploading file to storage (random size, random time);
  - remote file existence check;
  - file removal after some time.

To produce such a load on the deployed testbeds, a set of tools supporting most significant LRMS and MSS used in the project has been developed. Currently it supports XRootD, Condor, Torque/PBS and OGE. It provides a possibility to tune such event parameters as start time, file size, job memory, CPU usage, etc. Event parameters specified have an uniform statistical distribution by default, and maximum values could be adjusted to the cluster's configuration.

The load test suite is written in Python and Bash and provides a command line interface. Event series generation is started by cron system scheduler.

Conclusion

The JINR testbed for ATLAS Tier-3 sites monitoring tools development and testing is successfully deployed and used. It was used for installation and testing of numerous flavours of software environments widespread on ATLAS Tier-3 sites. A set of Ganglia-based monitoring tools for the most popular LRMS and MSS used on Tier-3 sites have been developed and tested on the testbed’s basis. The created infrastructure and methods are planned to be used in other projects like global data transfers monitoring in the nearest future.

References

[1] http://root.cern.ch/drupal/content/proof

[2] http://www.adaptivecomputing.com/products/open-source/torque/

[3] http://research.cs.wisc.edu/condor/

[4] http://www.oracle.com/us/products/tools/oracle-grid-engine-075549.html

[5] http://www.platform.com/workload-management/high-performance-computing/lp

[6] http://xrootd.slac.stanford.edu/

[7] http://www.dcache.org/

[8] https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm

[9] http://nfs.sourceforge.net/

[10] http://www-03.ibm.com/systems/software/gpfs/index.html

[11] http://www.lustrefilesystem.com/

[12] http://openvz.org

[13] http://xen.org

[14] http://gridedu.jinr.ru

[15] Artem Petrosyan, Danila Oleynik, Sergey Belov, Julia Andreeva, Ivan Kadochnikov “ATLAS off-Grid sites (Tier 3) monitoring. From local fabric monitoring to global overview of the VO computing activities”, to appear in Proceedings of CHEP2012 conference, New York, USA, May 21– 25, 2012.