ATLAS off-Grid sites (Tier-3) monitoring. From local fabric monitoring to global overview of the VO computing activities

Artem Petrosyan[1], Danila Oleynik[1], Sergey Belov[1], Julia Andreeva[2], Ivan Kadochnikov[1] on behalf of the ATLAS Collaboration

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

[2]CERN IT/ES, CH-1211 Geneva 23, Switzerland

Abstract. ATLAS is a particle physics experiment on Large Hadron Collider at CERN. The experiment produces petabytes of data every year. The ATLAS Computing model embraces the Grid paradigm and originally included three levels of computing centers to be able to operate such large volume of data. With the formation of small computing centers, usually based at universities, the model was expanded to include them as Tier-3 sites. The experiment supplies all necessary software to operate typical Grid-site, but Tier-3 sites do not support Grid services of the experiment or support them partially. Tier-3 centers comprise a range of architectures and many do not possess Grid middleware, thus, monitoring of storage and analysis software used on Tier-2 sites becomes unavailable for Tier-3 site system administrator, therefore Tier-3 sites activity becomes unavailable for virtual organization of the experiment. In this paper ATLAS off-Grid sites monitoring software suite is presented. The software suite enables monitoring on sites, not covered by ATLAS Distributed Computing software.

1. Introduction

The ATLAS Distributed Computing activities concentrated so far in the "central" part of the computing system of the experiment, namely the first 3 tiers (CERN Tier-0, the 10 Tier-1s centers and about 50 Tier-2s). This is a coherent system to perform data processing and management on a global scale and host (re)processing, simulation activities down to group and user analysis.

Many ATLAS Institutes and National Communities built (or have plans to build) Tier-3 facilities. The definition of Tier-3 concept has been outlined. Tier-3 centers consist of non-pledged resources mostly dedicated for the data analysis by the geographically close or local scientific groups.

Tier-3 sites comprise a range of architectures and many do not possess Grid middleware, which would render application of Tier-2 monitoring systems useless [1].

2. Tier-3 task force

In March 2011 a proposal was approved, which describes a strategy of development monitoring software for non-pledged resources: Tier-3 Monitoring Software Suite (T3MON) [2]. T3MON software package should meet the requirements of the ATLAS collaboration for global monitoring of ATLAS activities at Tier-3 sites, and the needs of Tier-3 site administrators. The solutions implemented in frames of this project are expected to be generic, so other Virtual Organizations (VO), within or outside of LHC experiments, can use them.

Software suite should satisfy the following:

- allow a site administrator to monitor local Tier-3 fabric(s);
- provide a global monitoring view to the services provided by the Tier-3 center, namely:
    - data transfers to the site and between sites;
    - data processing and analysis.

Main components of the system are:

- a software suite for local site monitoring – "T3MON-SITE";
- information system which should aggregate and visualize data from distributed Tier-3 sites at a global VO level – "T3MON-GLOBAL".

The main requirements for "T3MON-SITE" package are simple installation and support, intuitive user interface. The package should provide a low level monitoring of all site resources, status and performance of the hardware components, activities of the VO at the site. The toolkit should also include monitoring of data files located at the site. The toolkit should foresee a possibility for propagation of the aggregated monitoring metrics of the VO activities at the site to the VO central Tier-3 monitoring system ("T3MON-GLOBAL").

Central Tier-3 monitoring should be based on data collected from the local monitoring systems at Tier-3 sites. This data contain aggregated monitoring metrics of VO job processing and data transfer at a given Tier-3 site. The service must be scalable and has a minimal impact on the local resources. The set of the monitoring metrics as well as its granularity have to be defined by ATLAS.

3. Implementation of "T3MON-SITE"

In the light of the results of Tier-3 survey and in accordance with the requirements, we developed a package based on the Ganglia [3] monitoring system. Ganglia is an open-source package used for real-time monitoring of large UNIX clusters. Each node in a Ganglia system runs a daemon that reports on the state of its host in the form of performance metrics including memory, CPU, load, disk and network statistics. Collectors gather data produced by the daemons and store it in round-robin database. The information is typically presented in the form of plots via a web-server, but can be also obtained in XML format and consumed by various clients.

The main development effort was concentrated on enabling plug-ins for PROOF [4] and XRootD [5] monitoring through Ganglia. PROOF, The Parallel ROOT Facility, is an extension of ROOT (a framework for data processing) intended to parallelize certain class of tasks and could be considered as an alternative to batch systems for physics analysis purposes. XRootD is a highly scalable architecture and services for data access; it is widely used for distributed data handling and federation.

The PROOF plug-in contains a job accounting database, which is used to provide Ganglia with status information and send hourly messages containing data on file and job statistics, to the Dashboard [6], the system for monitoring of distributed computing systems of the LHC virtual organizations. In this case ActiveMQ [7] acting as message broker.

The XRootD monitoring makes use of both the summary and detailed XRootD monitoring streams produced by the XRootD servers. The monitoring daemons receive monitoring data as UDP packets, and after processing the information is displayed in Ganglia and sent through ActiveMQ. Figure 1 shows T3MON-SITE dataflow.

The methods and framework developed for implementing the PROOF and XRootD plug-ins will facilitate further development of plug-ins for monitoring of other software on Tier-3 sites, namely workload management solutions Condor [8], Torque [9], OGE [10] and distributed file system Lustre [11].

Condor plug-in utilizes the database of the Quill, the operational data logging system for Condor, to provide monitoring data for Ganglia and Dashboard. Quill is a natural part of Condor and gives enough information to monitor Condor daemons and queues as well as job statistics.

PBS/Torque queue and server status is obtained using the PBSQuery Python library, while job information is parsed directly from PBS accounting logs.

Lustre monitoring plug-in is based on reading the information provided in the virtual filesystem */proc/fs/lustre* and reporting it to Ganglia.

Job information from Condor and PBS/Torque is sent upward to Dashboard through ActiveMQ, while both status and job information is displayed in Ganglia.

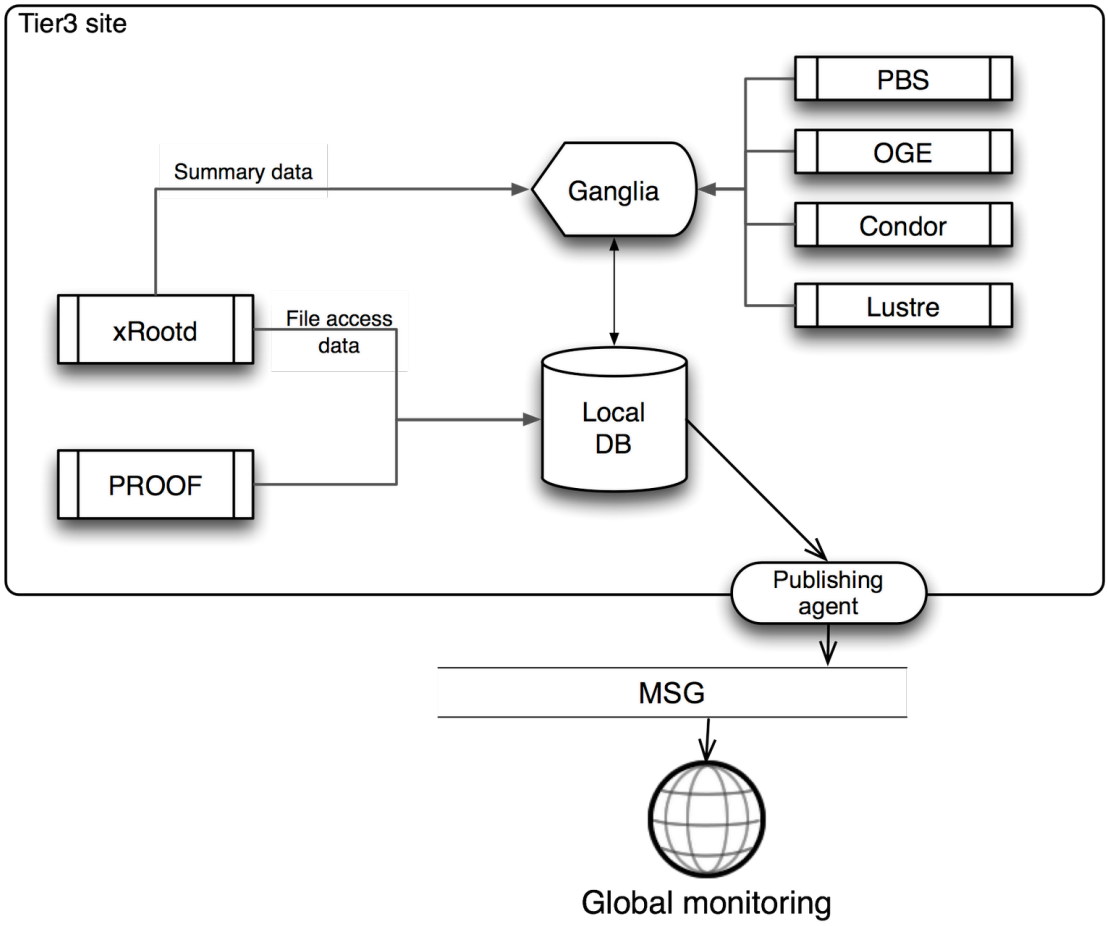

Figure 1. T3MON-SITE dataflow.

4. Implementation of "T3MON-GLOBAL"

The central Tier-3 monitoring system is based on monitoring data published by Tier-3 sites and should provide a global picture of how ATLAS uses Tier-3 resources. The necessary condition for the development of the central Tier-3 monitoring system is consistent registration of the Tier-3 sites in the ATLAS Grid Information System (AGIS) [12]. Another important factor is encouraging the Tier-3 user community to use data transfer and job submission systems which are instrumented for reporting the monitoring data, for example Ganga, Athena, DDM DQ2.

The system consists of several components (see Figure 2):

- Publishing agents, which run at Tier-3 sites, interact with the local monitoring systems, aggregate and publish monitoring metrics to the message bus. As a transport layer, we use the Apache ActiveMQ messaging system. Apache ActiveMQ is an open source messaging system which was recently evaluated as a standard messaging solution for the WLCG infrastructure.
- Data collector receives information through ActiveMQ message broker. Collected data is being recorded in the central data repository (based on HBase, the Hadoop database [13]).

- Data presentation layer includes an interactive user interface and an API for data export. Tier-3 views will be enabled in the existing Dashboard ATLAS monitoring systems, namely ATLAS DDM Dashboard, ATLAS Global job monitoring and Dashboard Transfer monitoring.

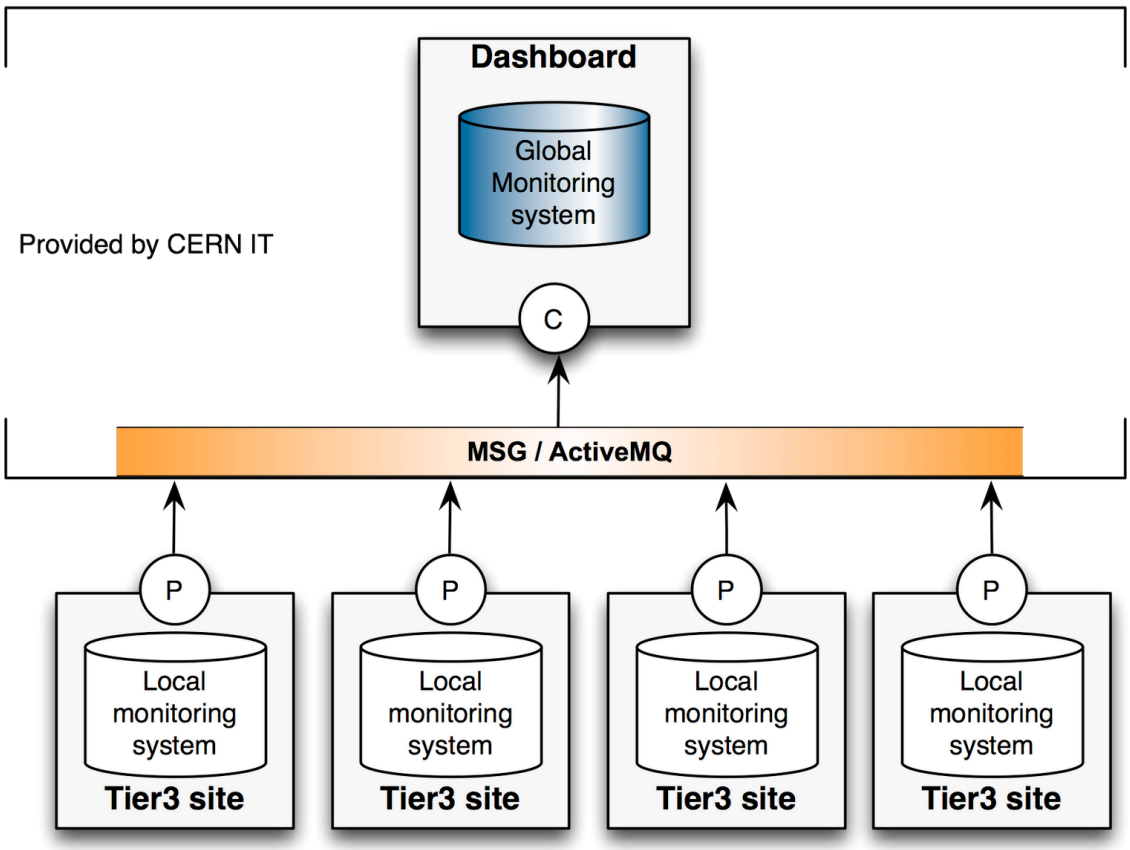

Figure 2. T3MON-GLOBAL logical scheme.

5. Project infrastructure

There are two main goals for the infrastructure of T3MON project. The first one is to deliver monitoring packages' distribution to site administrators in the most clear and convenient way. The second one is to give to the developers means to manage code base, distributions, documentation, and product testing procedures and have a feedback from end users.

T3MON monitoring tools are distributed as RPM [14] packages via special YUM [15] repository. These packages are built and tested for Scientific Linux 5 and Python 2.6 as a main target software platform. To install required packages, site system administrator has just to set up few YUM repositories configuration files (for T3MON stable and externals repositories, standard EPEL and DAG) and install desired tools from T3MON. Completely the same way is used by developers on the testbed; packages to be tested are taken also from testing and nightly repositories. All the project repositories are rebuilt once a day to provide fresh packages versions for the corresponding SVN packages releases.

Packages build and code release and versioning system is based on the tools developed within Dashboard project [6]. Originally based on standard tools as Python's *distutils* package and SVN versioning and revision control system, Dashboard's code management and build system significantly extended their functionality and flexibility towards better version management, packages build and deployment. In T3MON project this system was slightly improved (support for custom Python

version, remotely located external packages, extension of build scripts flexibility, and some other technical points).

6. JINR Tier-3 testbed

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

- validation of existing monitoring tools for each of the used component;
- development and debugging of new monitoring tools.

The activities listed above imply the following:

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

For estimating this aims, a testbed based on VM technology was deployed in JINR. PBS/Torque, Lustre, Condor, XRootD and PROOF clusters were installed. Testbed is based on virtualization technology, at the moment there are 18 virtual machines, all run on one physical server (AMD Athlon 64x2 Dual Core 3800+, 4Gb RAM, 320Gb HDD). Test load suite provides:

- job events;
- random submissions with configurable frequency;
- adjustable memory usage;
- CPU load;
- file events;
- uploading file to storage (random size, random time);
- remote file existence check;
- deletion of the file after configured period of time.

Ganglia packages were installed at each cluster; all clusters are being shown via one main web interface [16].

7. Conclusion

Monitoring tools developed within T3MON project allow having information on Tier-3 sites operation both on local and global levels. Most popular batch systems and mass storage systems used on real Tier-3 sites are supported.

On the level of site, there are several features quite useful for site's administration. There are detailed monitoring of the local fabric (overall cluster or clusters monitoring, monitoring each individual node in the cluster, network utilization). Job processing is monitored based on information from batch system. For of the mass storage system created tools make possible to watch such significant parameters as total and available disk space, number of connections, I/O performance.

On the global level, T3MON provides file transfers information and jobs statistics to the Dashboard. It permits having a view of general operation of such heterogeneous non-pledged and even off-grid resources as Tier-3 sites are.

T3MON software was installed and tested on volunteer sites such as Tier-3 sites at JINR, BNL and on the test site of Kyiv Polytechnic Institute in Ukraine.

Installation and configuration instructions can be found at project's home page [17].

References

[1] R. Brock et al., U.S. ATLAS Tier 3 Task Force, Preprint U.S. ATLAS, 2009.

[2] J. Andreeva et al., Tier-3 Monitoring Software Suite (T3MON) proposal, ATLAS note, 2011.

[3] Ganglia monitoring system, http://ganglia.sourceforge.net/

[4] PROOF, Parallel ROOT Facility, http://root.cern.ch/drupal/content/proof

[5] XRootD, http://xrootd.org/

[6] J. Andreeva et al., Experiment Dashboard for Monitoring of the LHC Distributed Computing Systems, J. Phys.: Conf. Ser. 331 (2011) 072001.
[7] ActiveMQ, message broker, http://activemq.apache.org/
[8] D. Thain et al., Distributed Computing in Practice: The Condor Experience, Concurrency and Computation: Practice and Experience, Vol. 17, No. 2-4, pp. 323-356, Feb-Apr, 2005.
[9] Torque resource manager, http://www.adaptivecomputing.com/products/open-source/torque/
[10] Oracle Grid Engine, http://www.oracle.com/us/products/tools/oracle-grid-engine-075549.html
[11] Lustre distributed file system, http://lustre.org
[12] A. Anisenkov et al., ATLAS Grid Information System, to appear in Proceedings of CHEP2012 conference, New York, USA, May 21– 25, 2012.
[13] HBase, the Hadoop database, http://hbase.apache.org/
[14] RPM package manager, http://www.rpm.org/
[15] YUM software package manager, http://yum.baseurl.org/
[16] Monitoring web interface example on testbed, http://vm01.jinr.ru/ganglia/
[17] T3MON project home, https://svnweb.cern.ch/trac/t3mon/wiki/T3MONHome