Automating ATLAS Computing Operations using the Site Status Board

Andreeva J.[1], Borrego Iglesias C.[2], Campana S.[1], Di Girolamo A.[1], Espinal Curull X.[3], Gayazov S.[4], Magradze E.[5], Nowotka M.M.[6], Rinaldi L.[7], Saiz P.[1], Schovancova J.[8], Stewart G.A.[1], Wright M.[9], for the ATLAS Collaboration

[1]CERN, European Organization for Nuclear Research, Geneva, Switzerland
[2]Institut de Física d'Altes Energies (IFAE), Universitat Autònoma de Barcelona, Edifici Cn, ES-08193 Bellaterra (Barcelona) Spain
[3]Port d'Informació Científica (PIC), Universitat Autònoma de Barcelona, Edifici D, ES-08193 Bellaterra (Barcelona) Spain
[4]Budker Institute of Nuclear Physics, 11, akademika Lavrentieva prospect, Novosibirsk, 630090 Russia
[5]II. Physikalisches Institut, Georg-August Universität Göttingen, Friedrich-Hund-Platz 1, D-37077, Göttingen, Germany
[6]Warsaw University of Technology, Pl. Politechniki 1, 00-661 Warsaw, Poland
[7]Instituto Nazionale Fisica Nucleare, Viale Berti Pichat 6/2, Bologna I-40127, Italy
[8]Institute of Physics, Academy of Sciences of the Czech Republic, Na Slovance 2, CZ-18221 Prague 8, Czech Republic
[9]School of Physics and Astronomy, Kelvin Building, University of Glasgow, Glasgow G12 8QQ, Scotland (United Kingdom)

E-mail: `atlas-adc-ssb-devs@cern.ch`

ATL-SOFT-PROC-2012-048
22 May 2012

Abstract. The automation of operations is essential to reduce manpower costs and improve the reliability of the system. The Site Status Board (SSB) is a framework which allows Virtual Organizations to monitor their computing activities at distributed sites and to evaluate site performance. The ATLAS experiment intensively uses the SSB for the distributed computing shifts, for estimating data processing and data transfer efficiencies at a particular site, and for implementing automatic exclusion of sites from computing activities, in case of potential problems. The ATLAS SSB provides a real-time aggregated monitoring view and keeps the history of the monitoring metrics. Based on this history, usability of a site from the perspective of ATLAS is calculated. The presentation will describe how the SSB is integrated in the ATLAS operations and computing infrastructure and will cover implementation details of the ATLAS SSB sensors and alarm system, based on the information in the SSB. It will demonstrate the positive impact of the use of the SSB on the overall performance of ATLAS computing activities and will overview future plans.

1. Introduction

The Worldwide LHC Computing Grid (WLCG) collaboration provides grid infrastructures spread among more than 150 computing centers (sites) in 35 countries [1] for handling large amounts of data produced by the High Energy Physics experiments at the Large Hadron Collider (LHC) at CERN [1]. The distributed infrastructure guarantees access to the data for more than

8000 physicists involved in LHC collaborations, regardless of their physical location. The grid computing infrastructure of the ATLAS experiment is a part of WLCG and is dedicated to maintenance, distribution and analysis of the ATLAS experimental data, so far there were registered about 10 PB of data at Tier 0 only in terms of the ATLAS experiment.

Considering the large amount of hardware resources and the heterogeneous infrastructure involved in the ATLAS grid computing it is challenging to provide reliable and stable performance. The response to the challenge is an effective monitoring and real-time action-taking (operations) carried out by ATLAS Distributed Computing (ADC) shifters and experts [2] as well as by site administrators and clouds squads [3]. Despite the efficient monitoring and the manual action-taking, issues remain like delays in the problem detection or action-taking process. These issues cause degradations of ADC service quality and efficiency. In order to avoid this it is necessary to reduce manual interactions by the introduction of automatic action-taking tools.

The development of such tools is one of the main goals of ADC operations. The automatic action-taking framework described in this work was developed using the ATLAS Site Status Board (SSB) [4], as described in Sec. 2. the SSB focuses on the collection, aggregation and publication of monitoring information about ADC critical activities. It also includes the ATLAS Site Topology [5] and an automatic alert system. Based on information provided by the SSB it is possible to define criteria for alerts or actions, allowing to interpret existing policies and instructions, defined by ADC operations, in terms of the SSB metrics that enables the development of generic automatic exclusion/recovery tools, which is meant for site or its component exclusion or re-inclusion (recovery) from particular activities. In section 3 we describe two example of the SSB tool and its two implementations.

2. The SSB Framework

Monitoring is a key service for WLCG as well as for the ATLAS grid computing infrastructure. Based on monitoring information it is possible to measure the performance of services and to detect problems in order to resolve them in time. Evaluation of critical activities in the ATLAS grid computing determines the performance, the availability and the efficiency of the computational processes and facilities. Monitoring and management of these activities is performed by dedicated independent tools. The Site Status Board (see fig. 1), developed on top of the Dashboard Application project [6], aggregates the monitor information for critical activities and allows the visualization of the overall status of the sites on web pages in formats like html, graphs,csv,xml or json.

The aggregation of monitoring information is performed by the Collectors, which are scripts communicating and collecting data from several external monitoring sources. Collectors run periodically with configurable frequencies. The accumulated information is stored in predefined data structures that are called metrics. Metrics contain quantitative or qualitative characteristics of different services that are published via web applications and the links to the monitoring sources meant for getting more details. They allow users to evaluate ATLAS site activities, such as Data Transfer, Data Analysis and Data Processing. Metrics contain only current status information. The older information is stored in a database to trace the evaluation or degradation of particular site activities. Based on the information in the database, the SSB provides historical views for the ADC activities [7], [8], [9], [10].

The SSB framework provides several web views. Users can choose the appropriate view for spotting problematic sites or getting overall information about ongoing activities. The most convenient way for viewing the monitoring data is a table view [11], where each column corresponds to one metric. A row contains an overall status for one site. The criticality of information is highlighted by color codes and values of qualitative or quantitative characteristics. The notations used in the table are common for all metrics and allow the identification of

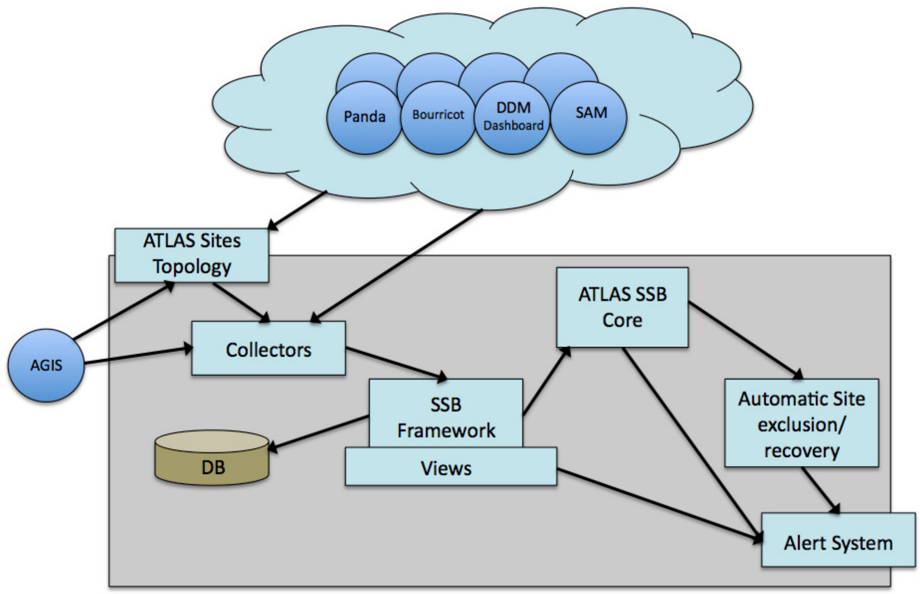

Figure 1. The workflow of the ATLAS SSB framework.

the problematic sites at once. In order to get details from source of the published values or characteristics, it is possible to click on and follow the link provided in the sufficient metrics.

In the process of information aggregation, one of the main challenges for the SSB framework was the mapping of the aggregated monitoring data with the information about sites provided by the ATLAS Grid Information System (AGIS) [12]. AGIS is based on information from the European Grid Infrastructure (EGI) topology and resources information system [13] (GOCDB) and the Open Source Grid (OSG) Information Management System (OIM) [14]. The SSB introduced the ATLAS Site Topology [5], which maps site attribute labels to the ones listed in AGIS. Thus all information provided by the SSB is consistent and coincides with official sources and naming conventions.

3. Exclusion/Recovery Framework and Alert System

The ability of the SSB framework to aggregate monitoring information from various sources and to combine them in a common data structure is a key feature that allows to easily identify the criticality level of given activities. The collected information is processed by inference rules based on which automatic actions might be triggered. Inference rules are sets of thresholds for metrics. The definition of these thresholds reflects existing policies, values from the ATLAS Computing Model [15], experiences of the ATLAS Distributed Computing and Operations Shifters (ADCoS) and experts as well as the SSB monitoring experience. The actions may interact with different kind of external frameworks. In sections 3.1 and 3.2 two implementations of such a system are described. The first one interacts with the Distributed Data Management (DDM) [16], and the second one with the Production and Distributed Analysis (PanDA) system [17]. The approach of these two implementations is transferable to other tools just by adding further sets of metrics, action scripts and inference rules.

In Figure 2 a sketch of the workflow of the main components of the automatic exclusion and recovery tools is shown.

Besides the execution of actions, the automatic exclusion/recovery framework interacts with the SSB alerts system. Via this system groups of people can be notified about predefined monitoring states. An example would be an automatic email to ADC shifters, experts or responsible personnel about taken actions.

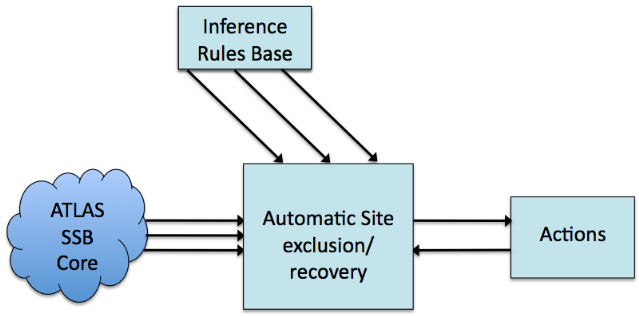

Figure 2. The automatic exclusion and recovery framework workflow.

3.1. ATLAS Site exclusion/recovery implementation for Data Analysis and Data Transfer

The purpose of this tool is the exclusion and re-inclusion(recovery) of sites from Data Analysis and Data Transfer activities. The ADCoS instructions and policies define precise criteria for the manual exclusion and recovery of sites from these activities. According to these criteria the SSB generates all needed metrics for automation of exclusion/recovery process. In Table 1, the actions and conditions for site exclusion/recovery from particular activity are summarized.

Activity	Reason	Action
Data Analysis	AFT_PanDA performance is lower than 50% in the last 24h	Exclude site
Data Transfer	DDMFT or SRMFT performance is lower than 60% in the last 48h	Exclude site
Data Transfer	DDMFT or SRMFT performance is 0% in the last 12h	Exclude site
All of the activities	DDMFT or SRMFT fails for more than three days	Exclude site
All of the activities	AFT_PanDA fails for more than five days	Exclude site
Data Transfer	DDMFT and SRMFT performance is greater than 60% in the last 24h	Recover site
Data Analysis	AFT_PanDA performance is greater than 60% in the last 24h	Recover site

Table 1. Exclusion/recovery details for Data Transfer and Data Analysis activities. (AFT - PanDA Analysis Queue Functional Tests; DDMFT - Data Distribution Managment Functional Tests; SRMFT - Storage Resource Management Functional Tests)

The conditions for automatic exclusion or recovery are tested hourly and each operation is reported to the subscribed site administrators and ADCoS by the SSB alert system.

3.2. ATLAS Site exclusion/recovery implementation for downtimes

The second SSB exclusion/recovery tool is meant for the automatic change of the PanDA queue status according to the downtime information. PanDA is the grid scheduling infrastructure to distribute ATLAS Monte-Carlo Simulation and data reprocessing jobs. All jobs are assigned to sites by submitting them to virtual PanDA site queues.

In Table 2 label of each column indicates the action that will be executed if there will be satisfied conditions listed in labels of the rows, e.g. if the queue status is “online” and the queue type is “Production” and there is a downtime for SRM in less than 12 hours or there is a downtime for computing element in less than 12 hours and there is no special comment for the

Actions/ Rules	Production Queue		Analysis Queue		
	set offline	recovery	set brokeroff	set offline	recovery
If queue status online	true	false	true	true	false
If queue status offline	false	true	false	false	true
If queue Status brokeroff	false	any	false	any	any
If queue type is Analysis	false	false	true	true	true
If queue type is Production	true	true	false	false	false
If SRM downtime in $6-12$ hours	true	true	false	false	false
If SRM downtime in $2-6$ hours	true	true	true	false	false
If SRM downtime in < 2 hours	true	true	false	true	false
If CE downtime in $0-12$ hours	true	true	any	any	false
If queue SRM is affected SRM	true	any	true	true	false
If queue CE is affected CE	true	any	any	any	false
If Comment is	empty	set.offline. by.SSB	empty	set.brokeroff. by.SSB	set.offline. by.SSB

Table 2. Rules and actions table for automatic exclusion/recovery due to downtimes.

queue, then the production queue will be switched offline. Recovery action means switching the queue in test mode with special comment HC.Test.Me that means notification to test the queue if it is ready to be included in production.

Setting a PanDA queue offline disables the job submission to the corresponding site. Sites may have several PanDA queues, and there are separate queues for production and analysis jobs. There are two types of downtimes: scheduled ones are declared in advanced by site administrators, and unscheduled ones are declared on short notice mainly due to unexpected site failures. During a site downtime different services can be affected. An outage of a Computing Element (CE) Service causes failures of jobs submitted via this service. An outage of the Storage Resource Management (SRM) Service will most likely cause failures of all jobs at a site. In order to avoid or minimize this it is important to take actions in time, i.e. to change PanDA queues accordingly. The automatic exclusion and recovery tool processes metrics with downtime information and metrics with PanDA queue information. The detailed criteria for automatic actions are given in Table 2. A production or analysis queue is set offline, set brokeroff (brokeroff is the status of the queue and indicates impossibility of sending new jobs to the queue. This status shows that the queue will go offline after all assigned jobs are finished and there will not be assigned new jobs to the queue) or recovered if all requirements in the corresponding column of Table 2 are fulfilled. The comment in the last row ensures that there is no interference with other queue manipulations done manually or by other systems.

The exclusion/recovery tool runs hourly and all actions are sent as notifications to sufficient egroups. Notifications are the emails sent by the SSB framework alert system [18]. More detailed information for tracking the taken actions and debug purposes is provided in log files.

4. Conclusions and Future Plans

The SSB provides a framework allowing the aggregation of all relevant information for automatic spotting of ADC activity or service degradation at a particular site. Generic solution of fetched monitoring information accuracy, instructions and policies defined by ADC led to automatic exclusion/recovery framework development. This tool is a step on the way of reaching ADC ultimate goal to reduce need for manual interactions in ATLAS Grid computing activities.

The generic structure of the automatic exclusion/recovery tool might be applied to other use cases. Once, there are defined metrics for criteria of site exclusion or recovery from particular activity, it is possible to write sufficient rules and action scripts. In order to allow further automatization of ADC operations the SSB plans to provide additional metrics to interpret more ADC policies in terms of inference rules that will lead to automation of more ADC operations and hence to reduction of failures in ATLAS Grid Infrastructure.

References

[1] WLCG Web Page `http://lcg.web.cern.ch/LCG`

[2] De K, Espinal X, Forti A, Korolkova E, Lehalm K, Love P, Schovancova J and Smirnov Y (ATLAS) 2011 *Journal of Physics: Conference Series 331 (2011) 072045*

[3] Schovancova J ATLAS Distributed Computing `http://nec2011.jinr.ru/docs/nec2011_Schovancova.pdf`

[4] SSB A ATLAS Site Status Board `http://dashb-ssb.cern.ch/ssb.html`

[5] SSB A ATLAS Site Topology `http://adc-ssb.cern.ch/SITE_EXCLUSION/ATLAS_sites.json`

[6] et al J A 2010 *Journal of Physics: Conference Series 219 (2010) 062003*

[7] Sonar View `http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview#currentView=Sonar&highlight=false`

[8] Netstat View `http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview#currentView=Netstat&highlight=false`

[9] SSB Reliability View `http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview#currentView=Reliability&highlight=false`

[10] SSB Usability View `http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview#currentView=Usability&highlight=false`

[11] SSB A Expandable Table Mode `http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview#currentView=default&highlight=false`

[12] ATLAS Grid Information System `https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasGridInformationSystem`

[13] GOCDB the official repository for storing and presenting EGI topology and resources information. `http://goc.egi.eu/`

[14] OSG Information Management System (OIM) `https://oim.grid.iu.edu/oim/hoame`

[15] Adams D, Barberis D, Bee C, Hawkings R, Jarp S, Jones R, Malon D, Poggioli L, Poulard G, Quarrie D and Wenaus T 2005 THE ATLAS COMPUTING MODEL `http://www.gridpp.ac.uk/eb/ComputingModels/atlas\_computing\_model.pdf`

[16] Branco M *et al.* 2008 *Journal of Physics: Conference Series 119 (2008) 062017*

[17] Maeno T 2008 *Journal of Physics: Conference Series 119 (2008) 062036*

[18] Borrego C, Di Girolamo A, Espinal X, Rinaldi L, Schovancova J, Andreeva J, Nowotka M M and Saiz P (ATLAS) 2011 *IBERGRID 2011 conference proceedings*