Integrated monitoring of the ATLAS online computing farm

S Ballestrero[1], F Brasolin[2], D Fazio[3], C Gament[3,4], C J Lee[5], D A Scannicchio[6], M S Twomey[7]

[1]University of Johannesburg Department of Physics, PO Box 524 Auckland Park 2006, South Africa
[2]INFN Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[3]CERN CH-1211 Genève 23, Switzerland
[4]Polytechnic University of Bucharest, Romania
[5]University of Cape Town, South Africa
[6]University of California, Irvine, CA 92697, USA
[7]University of Washington Department of Physics, Box 351560 Seattle WA 98195-1560, USA

E-mail: atlas-tdaq-sysadmins@cern.ch

Abstract. The online farm of the ATLAS experiment at the LHC, consisting of nearly 4100 PCs with various characteristics, provides configuration and control of the detector and performs the collection, processing, selection and conveyance of event data from the front-end electronics to mass storage.

The status and health of every host must be constantly monitored to ensure the correct and reliable operation of the whole online system. This is the first line of defense, which should not only promptly provide alerts in case of failure but, whenever possible, warn of impending issues.

The monitoring system should be able to check up to 100000 health parameters and provide alerts on a selected subset.

In this paper we present the implementation and validation of our new monitoring and alerting system based on Icinga 2 and Ganglia. We describe how the load distribution and high availability features of Icinga 2 allowed us to have a centralised but scalable system, with a configuration model that allows full flexibility while still guaranteeing complete farm coverage. Finally, we cover the integration of Icinga 2 with Ganglia and other data sources, such as SNMP for system information and IPMI for hardware health.

ATL-DAQ-PROC-2017-004
23/04/2017

1. Introduction

The online computing farm of the ATLAS [1] experiment at the LHC consists of nearly 4100 heterogeneous PCs. The status and health of every host must be constantly monitored to ensure the correct and reliable operation of the whole online system. The monitoring system is not critical for data taking, but it is our first line of defence: it promptly provides alerts in case of failure and warns of impending issues whenever possible.

The monitoring system is composed of Icinga 2 [2], for the active checks and alerting system, and Ganglia [3], as a data storage system and to provide the performance data useful for debugging and complementing the Icinga information. At the host level, the system is complemented by IPMI (Intelligent Platform Management Interface) [4] and SNMP (Simple

Network Management Protocol) [5] that are used to retrieve additional data directly from the host, respectively for hardware health and system information.

2. Icinga 2

Icinga is a fork of Nagios v3 [6]. In 2014 it replaced Nagios in the ATLAS monitoring [7], and is backward compatible: Nagios configurations, plugins and add-ons can all be used in Icinga. The main reason for updating the previous system was to simplify the setup by reducing the number of distributed Nagios servers ($\sim$ 80), while maintaining the same performance, and to increase the flexibility of the configuration.

At the end of 2015, Icinga was replaced by Icinga 2 to take advantage of the improved scalability and its native support for load balancing. Though Icinga/Icinga 2 retains all the existing features of its predecessor, it built on them to add many patches and features requested by the user community as described on its web site. Moreover, a very active community provides regular major releases, bug fixes and new features.

The monitoring system should be able to check up to 100000 health parameters and provide alerts on a selected subset. Two Icinga 2 servers are configured with the new built-in High Availability Cluster feature. The clustered instances assign an "active zone master", and this master writes to the Icinga database and manages configurations, notifications, and checks distribution for all hosts. Should the active zone master fail, the second server is automatically assigned this role. Furthermore, each instance carries a unique identifier to prevent conflicting database entries and "split-brain" behaviour.

Icinga 2 is currently successfully handling $\sim$ 4100 hosts and $\sim$ 83200 checks performed with different time intervals depending on the requirements. For each node $\sim$ 25 hardware and system parameters are monitored:

- Hardware
  - disk raid status
  - temperature
  - fan speed
  - power supplies
  - currents and voltages
  - firmware versions
  - amount of memory
- System Operation
  - cpu, memory and disk usage
  - configuration aspects such as kernel version and pending update
  - network: interface status, speed and bonding (if present)
  - services such as ntp, http/https, ssh, mailq

A notification system has been configured to warn the ATLAS Trigger and Data Acquisition (TDAQ) [8] System Administration team and the sub-detector experts who have requested it for their infrastructure. The system is configured to send emails of any server problems and hourly SMS for issues related to critical hosts or services.

A similar setup for the monitoring and notification has been deployed in the TDAQ TestBed: a test cluster configured similarly to the ATLAS TDAQ computing farm and designed for testing the TDAQ software. Here $\sim$ 400 hosts are monitored and $\sim$ 7100 checks are performed.

3. Ganglia

Ganglia is a software package designed for monitoring the workload and performance of multiple, large, possibly geographically distributed, high performance computing clusters. Unlike Icinga, it does not have advanced alerting capabilities.

A Ganglia server is used to monitor detailed information on CPU, disks and network performance: the data is saved as an RRD (Round-Robin Database [9]). The data are displayed via a Web User Interface, which, with its advanced functionalities, helps in performance analysis and forecasting.

Another Ganglia server is used as a data source for Icinga 2: additional information is retrieved from it instead of querying every single node; this reduces the number of active Icinga 2 checks performed by the server to the hosts.

An example of temperature and CPU usage plots from the Ganglia Web UI is shown in Figure 1: one can observe that the temperature (on the top) increases with the CPU usage (on the bottom) for the 40 worker hosts represented by the coloured lines.

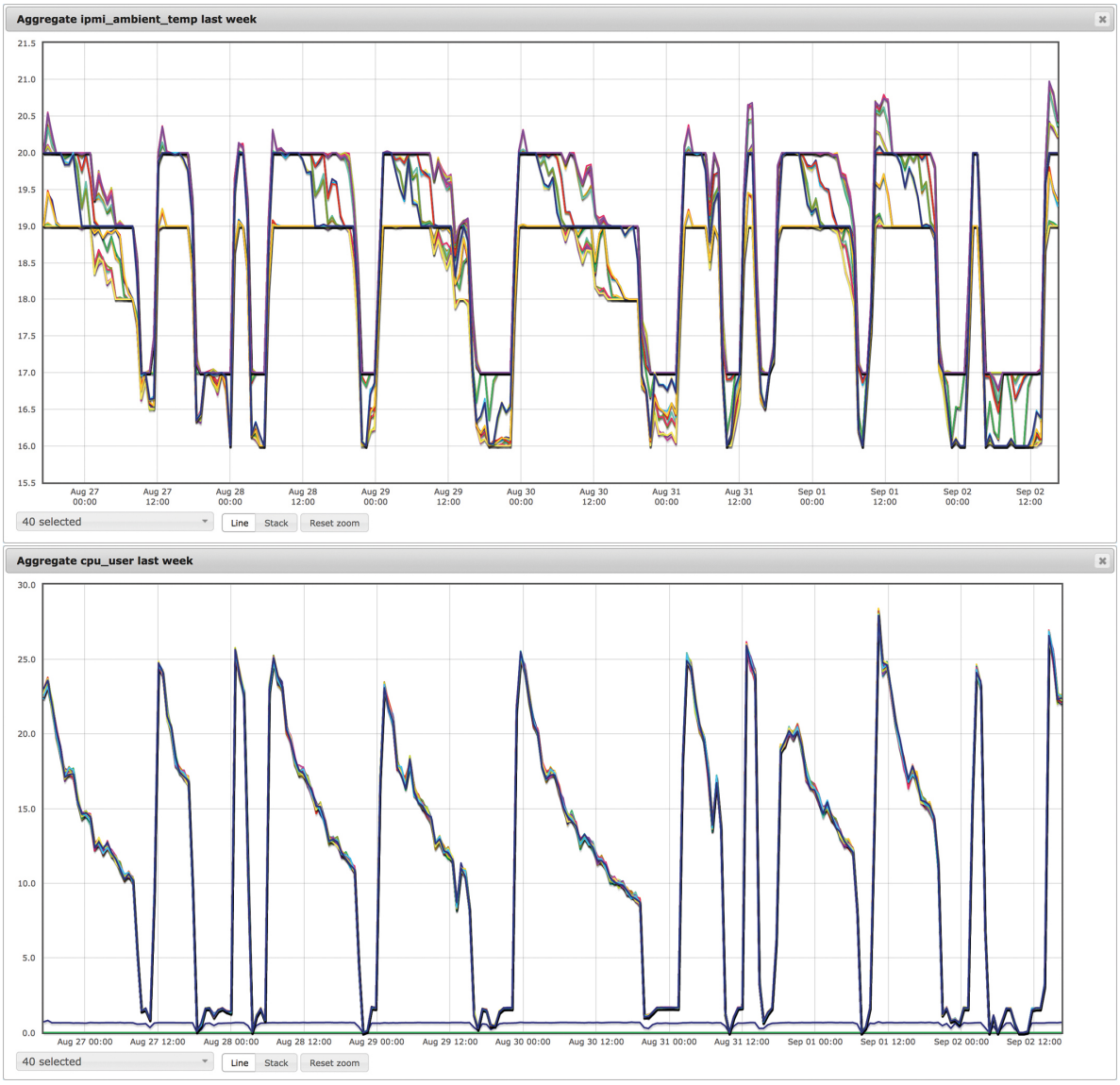


Figure 1. Ganglia temperature (top) and CPU usage (bottom) of 40 hosts in the same rack.

4. IPMI

A system has been put in place to monitor all the significant IPMI sensor information, which is provided by the Baseboard Management Controllers (BMC)[1] of the PCs in the farm. Generally 20 to 30 hardware sensors are monitored for each host. Amongst the most important are CPU and system temperatures, fan speeds, power supply status, various currents and voltages from the system.

Icinga 2 has to monitor multiple model types of machines supplied by different manufacturers; each of them interact differently with the different IPMI tools and IPMI v1.5/2.0 protocol specifications. Both FreeIPMI and IPMItool have been evaluated. FreeIPMI was better at interpreting the Original Equipment Manufacturer sensor's data among the different machines and provides more useful features.

The Icinga 2 server, using the FreeIPMI tools *ipmi-sensors* and *ipmi-sel*, queries the BMC of the hosts and fetches the entire output. This is then filtered (via *grep* and *sed*) for known System Event Log (SEL) entries and values of the sensors which are considered by Icinga as a correct state. The less concerning parts will appear as warnings and are filtered out as well. Whatever is left over is then considered to be a critical state.

5. SNMP

Simple Network Management Protocol (SNMP) is an Internet-standard protocol for collecting and organizing information about managed devices on IP networks. It is used in the Icinga 2 checks in two different ways.

A basic and standard check ("SNMP ping") is performed to check the responsiveness of a node by asking a simple SNMP question and waiting for the response.

The SNMP client is used to invoke a script on a machine just by sending an SNMP query. After the remote script finishes its standard/error output, return code and some other values (e.g. firmware version, error or warning messages, etc) are sent back to the client in an SNMP response. This script retrieves individual system information (e.g. network interfaces status and speed, firmware versions, services, amount of memory, disk availability, etc) that are available only on the machine itself.

For example, hardware RAID has been configured on the core servers and therefore it is important to get an immediate alert as soon as a problem appears. A complex and flexible script has been set up to retrieve the information about the hardware RAID status, taking into account the various hardware manufacturers monitored. The presence of different controllers required the use of different commands to check the status of volumes. The relative state of the charge level of the battery, if present, is also checked along with the cache policy.

6. Integrated monitoring system

The huge number of hosts to be monitored and the variety of configurations and settings required the creation of an automatic mechanism to produce the Icinga 2 configurations files. All the information related to the host hardware, operating system, and specific settings for the Icinga 2 checks, e.g. to override default thresholds or disabling a specific check, are stored in the ATLAS TDAQ System Administration in-house Configuration database (ConfDB) [10]. SQL queries and regular expressions, which exploit the adopted naming convention for the hosts, are used to extract data and select the right template for the host: this guarantees the full coverage of all the hosts, a per-type configuration and per-host tuning.

The configuration generator scripts, thanks to the improved and more powerful language provided by Icinga 2, leave us complete freedom to manually define the templates which will be applied: in this way we can modify and improve them without the need to intervene on

[1] The Baseboard Management Controller is described in the IPMI standard, see [4].

the generator code or on the database host definition. If needed we can add additional manual configurations which do not clash with the generated ones.

The Icinga 2 web interface provides an overview of every monitored host and the status of the checks. It allows us to have full control of the monitoring system: in the event of intervention, checks can be set to a downtime period or acknowledged in case of a temporary issue. Checks can also be forced to run again after the issue has been debugged and solved.

Icinga 2 retrieves data from each host in two different ways: in an active way where Icinga 2 queries either the BMC of the host in order to retrieve the IPMI SEL entries or through SNMP to run remotely a script and in a passive way where Icinga 2 queries the information from Ganglia. Each host, through the Ganglia monitoring daemon (*gmond*) service, sends to Ganglia the metrics of the various sensor values or the output of custom scripts.

The monitoring data flow related to a host is shown in Figure 2.

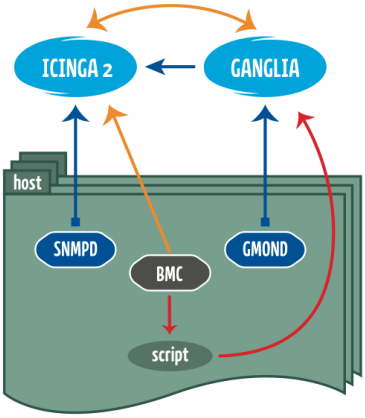


Figure 2. Host monitoring data flow.

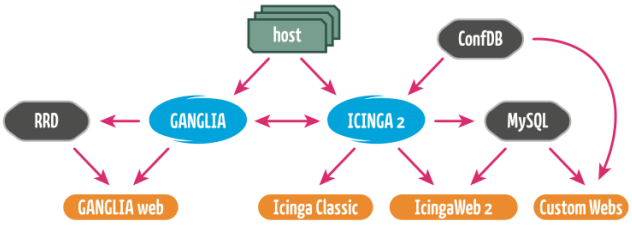


Figure 3. Overall monitoring data flow: from the host to the web interfaces.

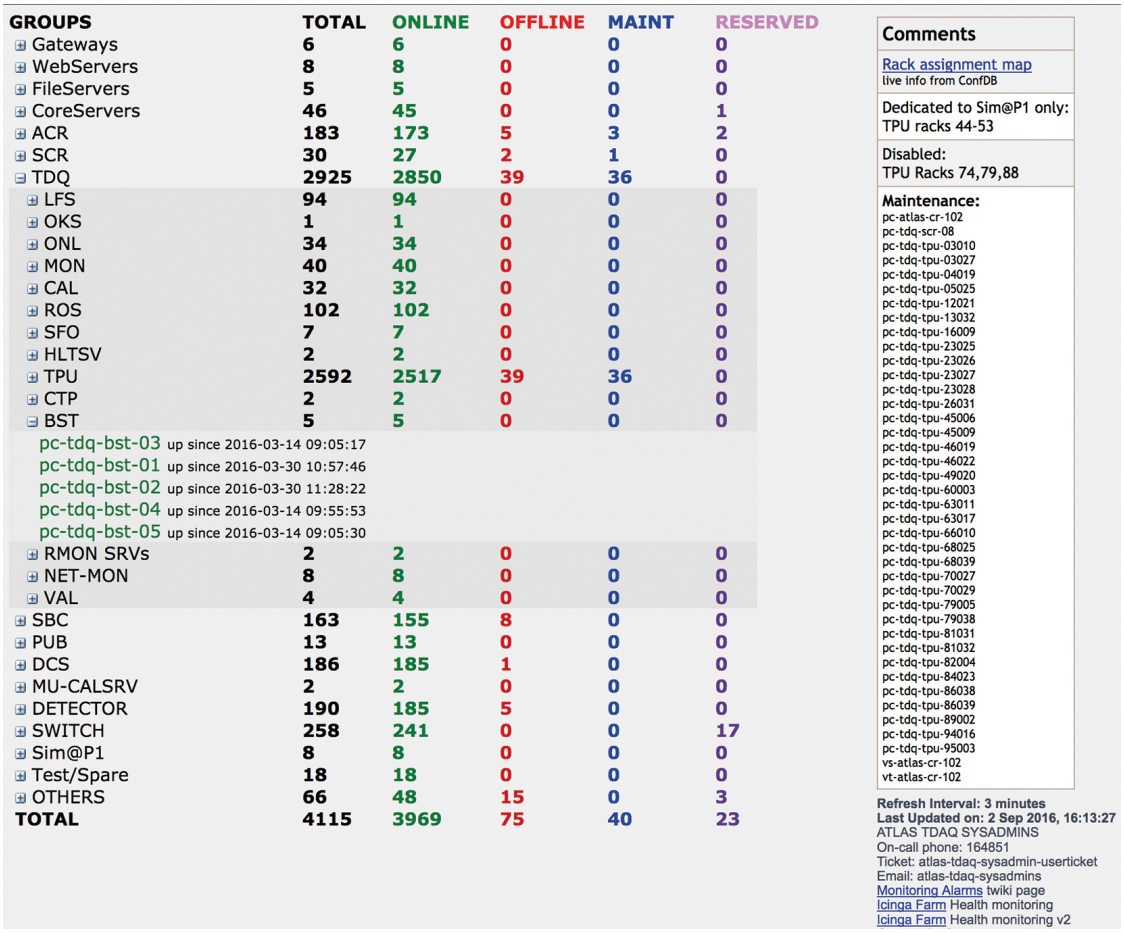
7. Farm overview

The data provided by the monitoring system is saved in a MySQL database to be used by various custom web interfaces (see Figure 3). They show, at a glance, the status of the farm: from the health status to the resources currently used by ATLAS operations.

A custom web page shows an overview of the operational status of each machine's group (Figure 4). This page makes it possible to condense some basic information in one single view, and quickly checking a particular group of hosts, instead of looking at various Icinga native pages. The web page was adapted from a previous version developed to group all the information which was spread among $\sim$ 80 Nagios servers. For example, it is possible to check how many machines are down or not operational, and how many are up and operational. It can be used as well to see if any of them is in under maintenance due to any ongoing intervention or reserved for being tested by the ATLAS TDAQ System Administration team, and therefore not available for the infrastructure.

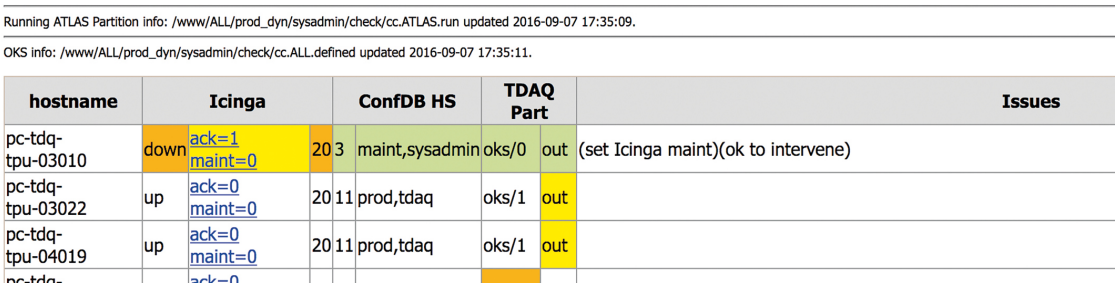
A cross consistency check custom web page has been developed to combine the overall information from Icinga 2, ConfDB and the ATLAS data taking status (Figure 5). ConfDB provides the status of a machine, for instance if it is in production or in maintenance due to any hardware or software intervention. The physical information of the machine, up or down, and if it has been acknowledged or set to downtime in the monitoring system, are retrieved from Icinga. The ATLAS data taking status of the machine is obtained from a dedicated database [12]: this informs the TDAQ System Administration team if the machine is currently in use for ATLAS

data taking, and therefore if it possible to intervene on that machine which needs maintenance, or if a machine, which has been fixed, is still not included in the ATLAS data taking and should be added back in.



GROUPS	TOTAL	ONLINE	OFFLINE	MAINT	RESERVED
Gateways	6	6	0	0	0
WebServers	8	8	0	0	0
FileServers	5	5	0	0	0
CoreServers	46	45	0	0	1
ACR	183	173	5	3	2
SCR	30	27	2	1	0
TDQ	2925	2850	39	36	0
LFS	94	94	0	0	0
OKS	1	1	0	0	0
ONL	34	34	0	0	0
MON	40	40	0	0	0
CAL	32	32	0	0	0
ROS	102	102	0	0	0
SFO	7	7	0	0	0
HLTSV	2	2	0	0	0
TPU	2592	2517	39	36	0
CTP	2	2	0	0	0
BST	5	5	0	0	0
pc-tdq-bst-03 up since 2016-03-14 09:05:17					
pc-tdq-bst-01 up since 2016-03-30 10:57:46					
pc-tdq-bst-02 up since 2016-03-30 11:28:22					
pc-tdq-bst-04 up since 2016-03-14 09:55:53					
pc-tdq-bst-05 up since 2016-03-14 09:05:30					
RMON SRVs	2	2	0	0	0
NET-MON	8	8	0	0	0
VAL	4	4	0	0	0
SBC	163	155	8	0	0
PUB	13	13	0	0	0
DCS	186	185	1	0	0
MU-CALSRV	2	2	0	0	0
DETECTOR	190	185	5	0	0
SWITCH	258	241	0	0	17
Sim@P1	8	8	0	0	0
Test/Spare	18	18	0	0	0
OTHERS	66	48	15	0	3
TOTAL	4115	3969	75	40	23

Figure 4. Status overview of the ATLAS computing infrastructure.



Running ATLAS Partition info: /www/ALL/prod_dyn/sysadmin/check/cc.ATLAS.run updated 2016-09-07 17:35:09.

OKS info: /www/ALL/prod_dyn/sysadmin/check/cc.ALL.defined updated 2016-09-07 17:35:11.

hostname	Icinga				ConfDB HS	TDAQ Part		Issues
pc-tdq-tpu-03010	down	ack=1 maint=0	20	3	maint,sysadmin	oks/0	out	(set Icinga maint)(ok to intervene)
pc-tdq-tpu-03022	up	ack=0 maint=0	20	11	prod,tdaq	oks/1	out	
pc-tdq-tpu-04019	up	ack=0 maint=0	20	11	prod,tdaq	oks/1	out	

Figure 5. Quick overview of the resources in use or not by ATLAS; "TDAQ Part" denotes the ATLAS data taking status.

8. Conclusions

The current monitoring and alerting system for the ATLAS online computing farm is based on Icinga 2 and Ganglia and provides the required information and alert notifications. It has proven its reliability and effectiveness in production, providing alerting and advanced performance monitoring. A big effort was made to automate and improve the checks and the notifications, and the generation of the Icinga 2 configuration files in order to simplify the system maintenance and to guarantee a whole coverage of the computing system.

The ATLAS Run 2 is anticipated to end in 2018 and there are no significant upgrades planned for the ATLAS online computing farm until then. This will allow us to evaluate the efficiency and performance of the current monitoring system and the possible needs in view of a future evolution.

References

[1] ATLAS Collaboration, The ATLAS Experiment at the CERN Large Hadron Collider, JINST 3 (2008) S08003
[2] Icinga: https://www.icinga.com/
[3] Ganglia: http://ganglia.sourceforge.net/
[4] Intelligent Platform Management Interface (IPMI): http://www.intel.com/design/servers/ipmi/
[5] Simple Network Management Protocol (SNMP): http://www.snmp.com/
[6] Nagios: http://nagios.org
[7] ATLAS TDAQ Sysadmin, Tools and strategies to monitor the ATLAS online computing farm, 2012 J. Phys.: Conf. Ser. 396 042053
[8] ATLAS Collaboration, ATLAS High Level Trigger, Data Acquisition and Controls: Technical Design Report, CERN/LHCC/2003-022 (2003), ISBN 92-9083-205-3
[9] RRDTool: http://oss.oetiker.ch/rrdtool/
[10] ATLAS TDAQ Sysadmin, Centralized configuration system for a large scale farm of network booted computers, 2012 J. Phys.: Conf. Ser. 396 042004
[11] ATLAS TDAQ Sysadmin, Upgrade and integration of the configuration and monitoring tools for the ATLAS Online farm, 2012 J. Phys.: Conf. Ser. 396 042005
[12] Dobson M et al., The ATLAS DAQ System Online Configurations Database Service Challenge, 2008 J. Phys.: Conf. Ser. 119(2008) 022004