Journal of Physics: Conference Series **396** (2012) 042004 doi:10.1088/1742-6596/396/4/042004

Centralized configuration system for a large scale farm of network booted computers

S Ballestrero[1,2], F Brasolin[3], G–L Dârlea[1,4], I Dumitru[4], D A Scannicchio[5], M S Twomey[1,6], M L Vâlsan[1] and A Zaytsev[7]

[1]CERN CH-1211 Genève 23, Switzerland
[2]University of Johannesburg, Department of Physics, PO Box 524 Auckland Park 2006, South Africa
[3]INFN Sezione di Bologna, Viale Berti Pichat 6/2, 40127 Bologna, Italy
[4]Politehnica University of Bucharest, Splaiul Independenţei 313, cod 060042, sector 6, Bucharest, Romania
[5]University of California, Irvine, CA 92697, USA
[6]University of Washington, Department of Physics, Box 351560 Seattle WA 98195-1560, USA
[7]Russian Acad. Sci., Siberian Div., Budker Inst. of Nuclear Physics (BINP), 11 Academician Lavrentiev prospect, Novosibirsk, 630090, Russia

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

Abstract. The ATLAS trigger and data acquisition online farm is composed of nearly 3,000 computing nodes, with various configurations, functions and requirements. Maintaining such a cluster is a big challenge from the computer administration point of view, thus various tools have been adopted by the System Administration team to help manage the farm efficiently. In particular, a custom central configuration system, ConfDBv2, was developed for the overall farm management. The majority of the systems are network booted, and are running an operating system image provided by a Local File Server (LFS) via the local area network (LAN). This method guarantees the uniformity of the system and allows, in case of issues, very fast recovery of the local disks which could be used as scratch area. It also provides greater flexibility as the nodes can be reconfigured and restarted with a different operating system in a very timely manner. A user-friendly web interface offers a quick overview of the current farm configuration and status, allowing changes to be applied on selected subsets or on the whole farm in an efficient and consistent manner. Also, various actions that would otherwise be time consuming and error prone can be quickly and safely executed. We describe the design, functionality and performance of this system and its web–based interface, including its integration with other CERN and ATLAS databases and with the monitoring infrastructure.

1. Introduction

The configuration system ConfDBv2 contains, in a MySQL database, information on all the computer systems of the ATLAS Trigger and DAQ (TDAQ) Online computing farm, and provides a flexible web interface for performing various operations. The tool aggregates specific system administration information with data from existing sources, such as the CERN central network database (LanDB), the ATLAS physical locations database (RackWizard [1]), etc. The automatic synchronisation with these external sources guarantees consistency across systems and decreases the likelihood of human error.

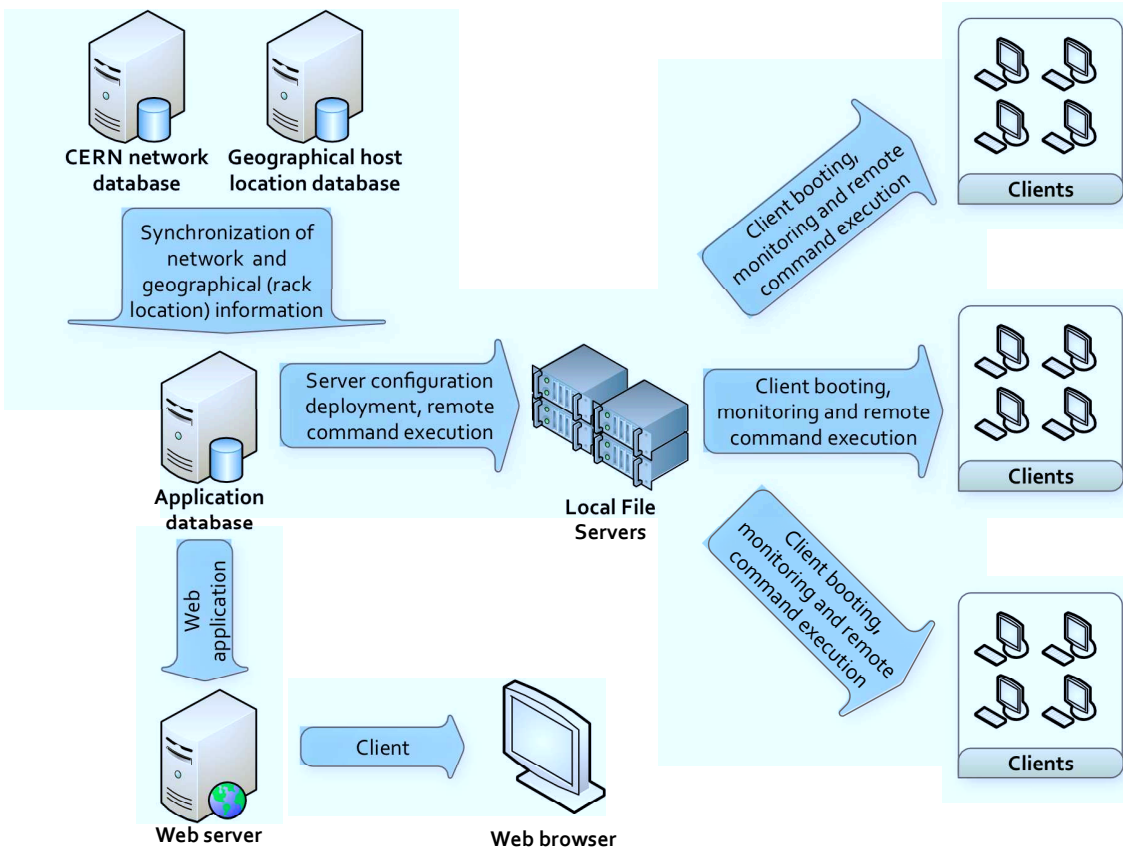

Figure 1. ConfDBv2 block diagram

The computer nodes registered in ConfDBv2 are grouped in different clusters according to their function and physical/network location. This results in having a compact view of a single cluster and being able to apply changes and deploy services on cluster–specific nodes, making the system more robust.

The diagram pictured in Figure 1 presents an overview of ConfDBv2 from two points of view: integration with the various external tools/databases on one hand and main functionalities which are available from the web graphical interface, on the other hand. The various blocks and links will be explained in the following sections.

2. ConfDBv2 main features

The ConfDBv2 Web Interface offers a large variety of functionalities and actions, out of which we will enumerate and briefly describe the most useful and significant ones.

2.1. DHCP configuration and management

The primary functionality provided by ConfDBv2 is the management of DHCP (Dynamic Host Configuration Protocol) and of PXE (Preboot eXecution Environment) booting. DHCP with static allocation (SHCP) is used for all the systems on the network, and PXE booting is used both for fully network booted hosts (usually clients/worker nodes) and for the initial OS installation of locally installed nodes (mostly servers and special purpose systems). For netboot clients,

the tool provides a complete set of management functionality; each client can be associated to one of the ∼90 boot servers (LFS, Local File Server), choosing not only a boot image to be loaded but also sets of kernel boot parameters, such as RAM disk size, IP source (generally DHCP), special locations for the boot images and post–boot scripts etc. Special operating system configurations are also available for nodes which need to boot special environments, for example for commissioning of new hardware or for diagnostics. The ConfDBv2 graphical interface allows to easily switch between the different operating systems available and to tweak kernel boot parameters according to the needs. A series of automatic actions (deployment of updated DHCP configurations, generation of PXE boot files) are triggered automatically with the change of these options, thus reducing the risk of inconsistencies.

The network booted nodes take advantage of the information stored in ConfDBv2 also during the automated configuration steps after booting, in particular for advanced configurations of network interfaces, such as bonding and VLANs.

2.2. Integration with the monitoring system

Another important functionality of ConfDBv2 is the management of the Nagios–based [2] monitoring and alerting system. ConfDBv2 stores the monitoring server for each particular node, but also the services and sensors which are being monitored. For greater flexibility sets of services and sensors to be monitored are grouped in templates which are then assigned to (sets of) hosts. Changing the monitoring server of a node or adding a node to the monitoring will automatically result in a deployment of the configuration for the monitoring service on the server. Additional information necessary for the correct functioning of the monitoring system is also embedded in the tool, such as alert policies, check intervals, notification targets, etc. The graphical interface allows changing these configurations in a consistent and fast manner.

2.3. Remote command execution

ConfDBv2 provides also a user–friendly interface for bulk command execution, with an easy interactive selection of the target systems based on various criteria, taking full advantage of the comprehensive information available for each node. The commands can be chosen from:

- a predefined set of IPMI (Intelligent Platform Management Interface) [3] requests
- a predefined set of shell commands (such as: reboot, shutdown, generate various configuration files); they are executed via SSH as superuser (root)
- custom shell commands
- a choice of operations specific to server systems

The command execution is performed in a fully parallel manner. The list of clients is split and distributed according to their respective boot server, and each server is responsible for executing the commands on its assigned clients. Where no boot server is available (locally installed nodes), the monitoring server is used instead for generating the remote command. For each server the execution of the command on the clients is also done in parallel, using up to 32 threads (matching the average number of clients handled by a server). A logging service is in place in order to ensure the traceability of the actions which are being performed.

2.4. Server–dedicated actions

Certain functions are targeted specifically to servers, namely the configuration deployment for DHCP and monitoring services. The DHCP daemon configuration deployment is automatically done at every DHCP configuration change on the affected server(s), but it can also be triggered manually. A similar mechanism has been implemented to handle changes which might occur on the monitoring configurations. The deployment of the monitoring configuration can also

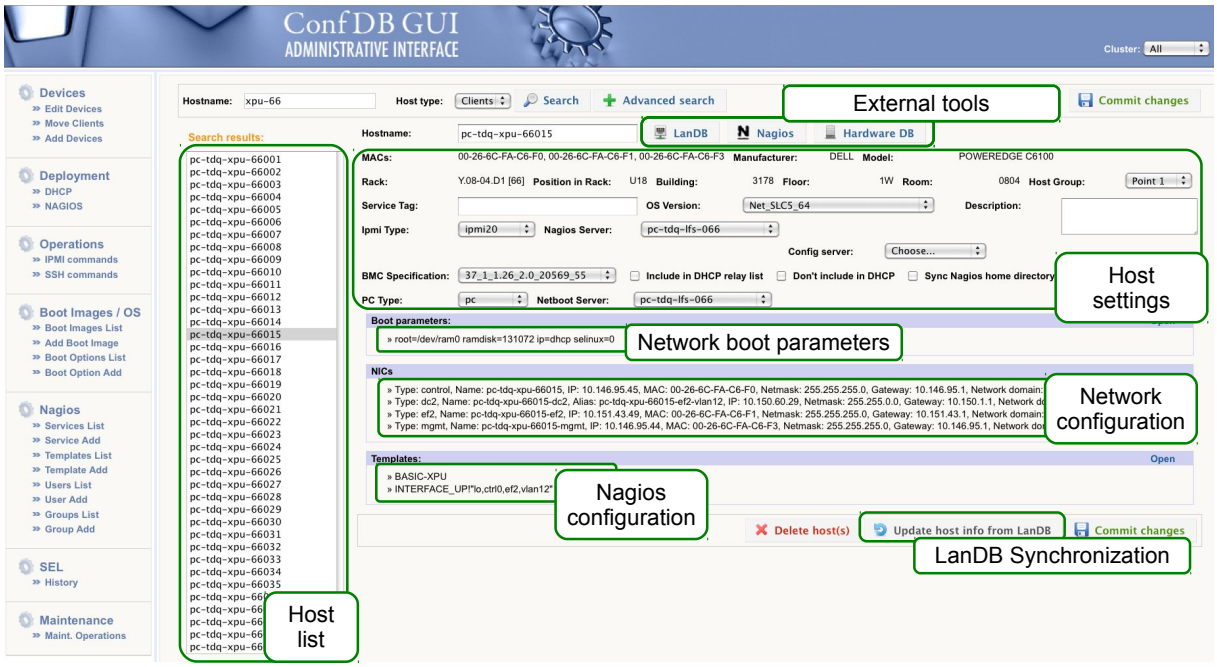

Figure 2. ConfDBv2 screenshot

be triggered manually on any selection of the monitoring servers. Removing, replacing or introducing a DHCP server requires the deployment of the DHCP relay configuration, which is also handled via ConfDBv2.

3. Integration with external tools

ConfDBv2 also integrates information on the functional status of each node via an internal hardware database. This includes the history of the hardware failures, with details of the affected components and of the actions taken for fixing them. The historical information has been used for analysing the maintenance effort and to validate the warranty needs and replacement plans. The current status of each node is also exported to the TDAQ infrastructure, allowing the automatic exclusion of malfunctioning nodes from the working system. Retired machines are marked as such, so that they are not accidentally included in the running farm; their history is kept in case they will be re-used in the future for other purposes. Finally, nodes can be marked as reserved, for maintenance or for various hardware or software validation tasks, thus remaining available for the intended tests solely.

The information on every device is being kept in sync with the central CERN networking database (LanDB). Direct access to the device description in LanDB is also available from the ConfDBv2 web interface. If more detailed information about the physical device is needed, there is the possibility of accessing its location and main hardware characteristics inside another ATLAS–specific database, RackWizard. RackWizard keeps information related to exact physical location inside the ATLAS server rooms, dimensions, power consumption characteristics etc.

4. Web–based user interface

A snapshot of the ConfDB graphical interface is given in Figure 2, with highlights on the most interesting features, that have been discussed in the previous sections. As it can be seen, the web interface offers an immediate and intuitive access to the main functionalities offered by the

tool.

Besides all the described functions, various actions which would otherwise be time consuming and error prone can be quickly and safely executed via the interface, such as:

- moving the boot clients from one server to another,
- changing the monitoring server,
- adding new devices in the ConfDBv2 database,
- adding and altering monitoring services and templates,
- adding boot options.

5. Conclusions

The ConfDBv2 centralized configuration system is a very flexible and robust tool which greatly improved the management of the ATLAS Trigger and Data Acquisition Online farm. Not only does it offer a quick overview of the current farm configuration and status, but it also allows changes to be applied on selected subsets or on the whole farm in an efficient and consistent manner. It also provides centralised access to all the main components of the computing system, thus acting as a unifying tool for most of the various utilities used to manage the ATLAS TDAQ computing farm.

References

[1] F. Glege, "The Rack Wizard, a graphical database interface for electronics configuration", 9th Workshop on Electronics for LHC Experiments, Amsterdam, The Netherlands, Sep–Oct 2003, pp.369–372, LECC 2003
[2] Nagios website: http://www.nagios.org
[3] Intelligent Platform Management Interface (IPMI); http://www.intel.com/design/servers/ipmi/