LHCb ONLINE INFRASTRUCTURE MONITORING TOOLS

L. Granado Cardoso, C. Gaspar, C. Haen, N. Neufeld, F. Varela, CERN, Geneva, Switzerland
D. Galli, Università di Bologna-INFN, Bologna, Italy

Abstract

The Online System of the LHCb experiment at CERN is composed of a very large number of PCs: around 1500 in a CPU farm for performing the High Level Trigger; around 170 for the control system, running the SCADA system - PVSS; and several others for performing data monitoring, reconstruction, storage, and infrastructure tasks, like databases, etc. Some PCs run Linux, some run Windows but all of them need to be remotely controlled and monitored to make sure they are correctly running and to be able, for example, to reboot them whenever necessary. A set of tools was developed in order to centrally monitor the status of all PCs and PVSS Projects needed to run the experiment: a Farm Monitoring and Control (FMC) tool, which provides the lower level access to the PCs, and a System Overview Tool (developed within the Joint Controls Project – JCOP), which provides a centralized interface to the FMC tool and adds PVSS project monitoring and control. The implementation of these tools has provided a reliable and efficient way to manage the system, both during normal operations as well as during shutdowns, upgrades or maintenance operations. This paper will present the particular implementation of this tool in the LHCb experiment and the benefits of its usage in a large scale heterogeneous system.

INTRODUCTION

Due to the large nature of the LHCb experiment, a high number of PCs is needed to reliably and efficiently operate it. At this scale it is fairly unreliable to monitor and maintain the needed operation parameters without the help of a centralized tool that provides a global overview of the current computer infrastructure status. Furthermore, the heterogeneous nature of the underlying computer infrastructure is an added difficulty for monitoring and control of the experiment computer infrastructure.

The System Overview set of tools, is, therefore, based on a layered approach for monitoring different sets of parameters of the PC infrastructure:

- Hardware Layer – The hardware infrastructure (PCs) where the applications run
- Operating System Layer – The software infrastructure/environment where the applications run
- PVSS Infrastructure Layer – The supervisory control software running the experiment infrastructure/environment
- Application Layer – Specific software needed for system control and operation (PVSS Managers, OPC Servers, etc.)

SYSTEM OVERVIEW SET OF TOOLS

To monitor the parameters for the different monitoring layers, a set of tools were developed, which are able, for each of the existing environments, to gather the relevant parameters:

- Monitoring and Control Servers (FMC Tools)
  - FMC Linux tools [1] – to gather and publish Linux computers data and control processes and IPMI status;
  - FMC Windows Server – to gather and publish Windows computers data;
- GUI (Graphical User Interface) and "aggregation" Tools – to provide a centralized interface for monitoring and control;

Monitoring and Control Servers

The Monitoring and Control Servers gather data from the monitored nodes and publish the monitored parameters via DIM (Distributed Information Management) [2]. These servers are implemented differently for each of the operating system in use in LHCb and can be divided according to the layer they monitor:

Hardware Layer

OS independent:

- ipmiSrv – IPMI Server, which runs in only one LHCb node; This server polls the IPMI status of the PCs in the LHCb network, via ipmitool and publishes this in a DIM Service, which are available for subscription from other tools. The information obtained for each PC is the power status of the PC, and information pertaining the PC sensors which are available via IPMI (temperatures, fan info, currents and voltages)
- vmSrv – which runs on one LHCb node which can act on the hypervisor for the LHCb Virtualization Infrastructure; This server publishes power information for the VMs (Virtual Machines) on the LHCb system and is also able to control the power status via the hypervisor.

OS dependent:
For Linux:

- memSrv – Memory Server, runs on each of the nodes; gathers current memory usage information on the PCs
- cpuinfoSrv – CPU Info Server, runs on each of the nodes; gathers cpu information for the PC (number of cores, speed and type of CPU)

- cpustatSrv – CPU Statistics Server, runs on each of the monitored nodes; gathers CPU usage statistics for the PC
- fsSrv – File System Monitor Server, runs on each of the monitored nodes; gathers info the mounted filesystems and its usage
- nifSrv – Network Interface Monitor Service; monitors network traffic, uptime and statistics for each of the PC Network Interface Cards

For Windows:

- FMC Windows Server - Memory Usage, CPU information, CPU Statistics, File System information, Network Interfaces

Operating System Layer

For Linux:

- osSrv – Operating System Server, runs on each of the nodes; This server gathers operating system and kernel information

For Windows:

- FMC Windows Server - monitors Operating System information, Memory Usage, CPU information, CPU Statistics, File System information, Network Interface information and Processes and Services information.

PVSS Infrastructure Layer

- PVSS pmon process - A process monitor agent linked to each PVSS Project that runs independently from it. This agent monitors and publishes the state of PVSS processes. It can also act on these processes (start/stop/reset)

Application Layer

For Linux:

- tmSrv – Task Manager Server, which runs on each of the nodes; gathers the data for the running processes on each node and is also able to start processes on these nodes.
- psSrv – Process Monitor Server, runs on each of the monitored nodes; gathers info on the running processes on each PC
- pcSrv – Process Controller Server; this server keeps a dynamically manageable list of applications up and running on the farm nodes, stopping and/or restarting them as defined
- logSrv – Log Server, which runs on each of the nodes; collects the logs from the several FMC Linux tools running on the PCs

For Windows:

- FMC Windows Server – monitors processes and services information.

All these servers publish their data via DIM services which are then available for subscriptions from other tools. The services published follow a naming convention, independent of the system where the servers are running.

Note that while the Linux Servers have to run on each of the nodes being monitored, the Windows FMC Server runs on only one machine which has network access to all the windows machines which need to be monitored and the list of machines to be monitored must be defined on start time.

GUI and "Aggregation" Tools

FwFMC

The JCOP Framework [3] FMC component is a PVSS component which subscribes to the DIM services and commands published by the FMC servers and provides a GUI for easy interaction, monitoring as well as an archiving infrastructure for this data.

This tool can be configured from a database which holds the nodes and PVSS projects information as well as from a configuration file. It can also be configured manually adding the nodes and PVSS projects to be monitored.

FwSystemOverview Tool

The FwSystemOverview Tool reutilizes the data subscribed from the FwFMC tool and presents the status of the monitored computers in synoptic panels, with easy control and monitoring possibilities. It is also possible with this tool to monitor single processes and assign alarms for the unexpected dying of these processes. Alarms can also be configured for usage levels of CPU and memory.

The FwSystemOverview tool also has support for monitoring and controlling of PVSS Projects and is able to monitor individual managers via calls to the respective PVSS project Process monitor manager. This allows for PVSS project control for all the running projects from a single place. It is also possible to search for particular managers on all the running projects and act upon them simultaneously.

The monitoring level for each device can be defined and it is possible to monitor only relevant information for each device

FSM (Finite State Machine) [4] objects are also available from this tool, in order to create FSM hierarchies that enable the possibility of grouping the nodes and projects according to desired logical groups. This enables the possibility of grouping and controlling globally nodes and PVSS projects with similar functions or characteristics

LHCb ARCHITECTURE

The LHCb Online System is composed of 1747 PCs, of which 55 machines have Windows and 1692 have Linux installed. Also, of all these machines 34 are virtual machines.

1470 of these machines belong to a CPU Farm to perform the High Level Trigger, 167 are controls machines running the PVSS SCADA system and 110 are for infrastructure support (DBs, storage), webservers, reconstruction or data monitoring

These machines are all connected in the same network and are all accessible to each other within the network.

One machine houses the PVSS part of the System Overview Tool (FwFMC, FwSystemOverview) and provides a DIM DNS Node to where all the running Monitoring and Control Servers publish their data. This machine also serves as IPMI Master.

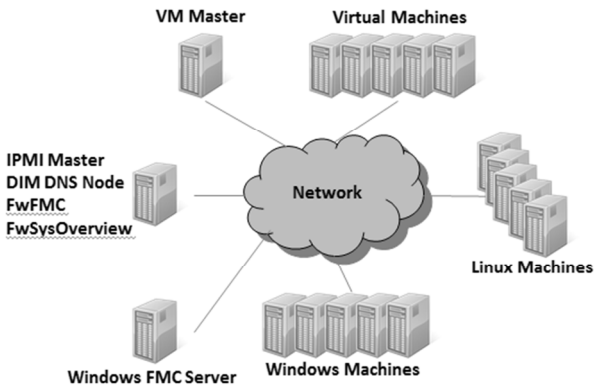

Figure 1: LHCb System Overview Architecture.

For the Hardware Layer monitoring, there are two machines that monitor the power status, fan status, temperatures and voltages:

- 1 IPMI Master – Runs the IPMI Server and polls the IPMI status from each of the nodes present on a configuration file; Controls the power of the configured nodes via IPMI commands
- 1 VM Master – Monitors and controls the power status of the Virtual Machines present on a configuration file.

Currently all of the available machines in the LHCb Online System are configured so as to be possible to monitor and control its power status.

Other parameters from the Hardware Layer (Memory and CPU usage monitoring) are configured on the nodes which are more prone to suffer more of high usage, either by running the respective servers if a Linux Machine or adding the node to a list of monitored nodes on the Windows FMC Server. Alarms are configured on these parameters so that eventual problematic situations are identified as soon as possible.

On the PVSS Infrastructure Layer, all PVSS projects have their status continuously monitored and all the individual managers of the projects are also monitored.

For the Application Layer, all the controls PCs and a few Linux PCs whose processes need to be monitored and controlled more attentively have also the task manager server running. Also the windows PCs which are configured in the Windows FMC server have the processes and services monitoring capability enabled.

LHCb USAGE

LHCb opted to develop the interface based on 2 hierarchies: one for monitoring and control of the hosts and one for monitoring and control of the PVSS projects. These hierarchies are divided according to sub-detector and function. This division allows for, at a quick glance of a synoptic panel, have an overview of the global system as well as the individual sub-detector or any particular function computer infrastructure.

This particular division also allows for the individual sub-detector to access only their specific monitored infrastructure and thus simplifies their particular needs for management.

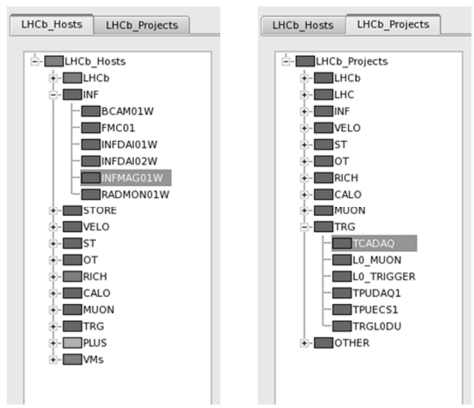

Figure 2: LHCb Hosts and Projects Hierarchies.

Using this division, it is easily identifiable on the "LHCb Hosts" hierarchy if there is any problem with the hardware of a particular system and its diagnose simplified - It is easy to identify the power status and memory usage of the machines as well as a list of running processes on it. It is common to send actions to the hosts (switch ON/OFF, reboot, kill processes) in order to fix the identified abnormalities.

Likewise looking at the "LHCb Projects" hierarchy gives you a global overview of the status of the PVSS SCADA control system and allows you the easy detection of any problem related with the PVSS projects which need to be running to ensure smooth and coherent operation of the experiment. It is commonly used to see if all the required PVSS projects are running and if any particular manager on any of the projects is functioning abnormally (process blocked) or not running (abnormally stopped, killed).

Another advantage of using the System Overview Tool is the possibility to search and globally manage the PVSS managers for all the configured PVSS projects (Fig.3). This allows an easier upgrade of the control software

available on the repositories as all the managers that would need to be stopped and restarted can now be managed from a single access place and in parallel.

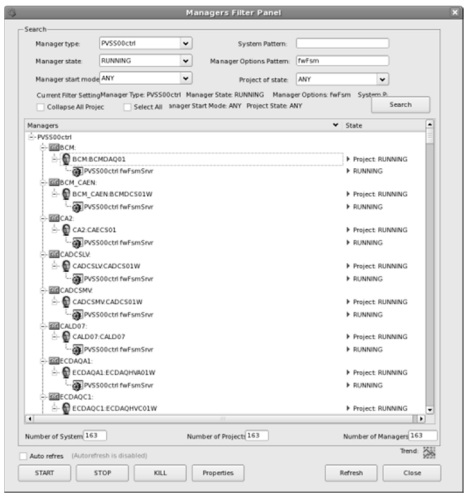

Figure 3: System Overview PVSS manager filtering.

CONCLUSION

The System Overview set of tools provide for a very elegant and user friendly central management tool, with many benefits to the administration of the LHCb online system. It enables an easy and complete monitoring system from a centralized place, permitting a global overview of the system status as well as a fine control over the several nodes, PVSS projects and processes of the system.

The system Overview tool is able to monitor an heterogeneous system with the same interface, detaching the monitoring and control interface from the monitoring and control servers which gather the data and are able to act on the nodes.

The system overview tool is also expandable to TCP enabled devices other than PCs, as soon as there is an available tool for these devices to publish their info into system overview.

REFERENCES

[1] F. Bonifazi et al., "The Monitoring and Control System of the LHCb Event Filter Farm", IEEE Transactions on Nuclear Science, Vol. 55, No.1, Feb. 2008.

[2] C. Gaspar, "DIM - A Distributed Information Management System for the Delphi experiment at CERN", IEEE Real Time Conference, 1993, Vancouver, Canada

[3] O. Holme et al., "The JCOP Framework", ICALEPCS 2005, Geneva, Switzerland.

[4] C. Gaspar and B. Franek, "Tools for the automation of large distributed control systems", IEEE Transactions on Nuclear Science., Vol. 53, NO. 3, June 2006

[5] Prozeßvisualisierungs - und Steuerungs system made by ETM Professional Control GmbH, Eisenstadt, Austria. http://www.pvss.com

[6] F. Varela, "Software management of the LHC Detector Control Systems", ICALEPCS, 2007, Knoxville, Tennessee, USA.

[7] "IPMI v2.0 specifications Document Revision 1.0", http://download.intel.com/design/servers/ipmi/IPMIv2_0rev1_0.pdf