MANAGEMENT OF THE LHCb NETWORK BASED ON SCADA SYSTEM

G. Liu*, University of Ferrara and INFN Ferrara Section, Ferrara, Italy and CERN, Geneva, Switzerland
N. Neufeld, CERN, Geneva, Switzerland

Abstract

LHCb employs two large networks based on Ethernet. One is a data network dedicated for data acquisition, the other one is a control network which connects all devices in LHCb. Sophisticated monitoring of both networks at all levels is essential for the successful operation of the experiment. The network management system is implemented based on a commercial SCADA system (PVSS II). We show here how a large scale network can be monitored and managed using tools originally made for industrial supervisory control.

LHCb-CONF-2009-047
02/11/2009

INTRODUCTION

The LHCb experiment [1] is one of the large particle physics experiments built on the Large Hadron Collider (LHC) at CERN. LHCb is dedicated to the study of the decays of B-hadrons produced at LHC in order to measure CP violation and search new physics. The LHCb detector is a single arm forward spectrometer, comprising several subdetectors and several global infrastructure systems. The online system in LHCb [2] comprises all aspects of experiment controls, data taking, Timing and Fast Control and the other online computing. It provides the IT (Information Technology) infrastructure for the entire experiment. Two large scale Ethernet networks are deployed for experiment control and data acquisition (DAQ), called control network and data network. The control network is a general purpose network, connecting the experiment control servers and all other devices based on Ethernet. It consists of two core control switches (Force10 E600) and ∼100 access switches (HP 2650). The data network is dedicated for DAQ. It is composed of one core DAQ switch and ∼50 access switches. The core DAQ switch is a high-end switch Force10 E1200i with 1260 Gigabit Ethernet (GbE) ports. These access switches are HP 3500 switches, which transport data from the core DAQ switch to the CPU farm. LHCb comprises two level triggers, the first level (L0) trigger and the high level trigger (HLT). L0 trigger is based on custom hardware, while HLT is based on the software running in a large CPU farm. The data network forwards the data selected by L0 trigger from readout boards to the HLT CPU farm. The L0 trigger rate is 1MHz, the average throughput is ∼35 GB/s. In HLT CPU farm, the event rate is further reduced to ∼2 KHz. The selected data are sent to local storage and then to CERN CASTOR tapes for permanent storage. The HLT algorithm needs all the fragments belonging to the same event, so the performance is quite critical and it is important not to lose any packet. Sophisticated monitoring of both networks at all levels is essential for the successful operation of the experiment.

In the following sections, the LHCb Experiment Control System (ECS) will be introduced, then the network monitoring system based on Supervisory Control and Data Acquisition (SCADA) system will be described in detail.

* gliu@cern.ch

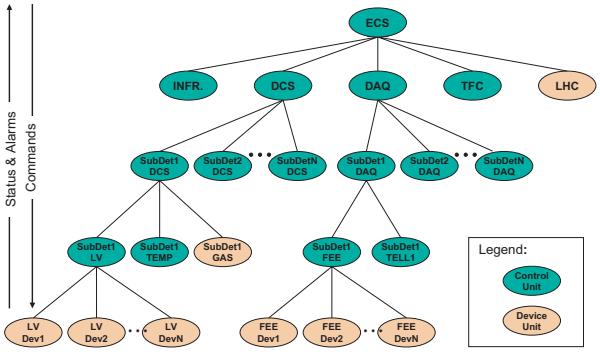

Figure 1: ECS architecture.

LHCB EXPERIMENT CONTROL SYSTEM

The LHCb ECS [3] handles the configuration, monitoring and operation of all experimental equipment involved in the different activities of the experiment. This encompasses not only the traditional detector control domains, such as high and low voltages, temperatures, gas flows, or pressures, but also the control and monitoring of the Trigger, Timing and Fast Control, and DAQ systems.

At CERN, the Joint Controls Project (JCOP) [4] was set up for all the LHC experiments. JCOP provides the framework and the tools to develop control systems. This framework is based on a SCADA system called PVSS II [5].

As shown in Fig. 1, LHCb ECS adopts a hierarchical structure to represent the structure of sub-detectors, subsystems and hardware components. The system comprises two types of nodes, Device Units (DU) which are capable of driving hardware and Control Units (CU) which can monitor and control the sub-tree. The State sequencing in the ECS system is achieved by using a Finite State Machine (FSM) package, based on SMI++ [6]. Device units and control units are all modeled as FSMs. The FSM represents the state of each sub-system, provides convenient mechanism to model the functionality and behavior of a component, and provides an intuitive user interfaces for experiment operators.

The network monitoring system is part of the experiment control system, which will be described in the following section.

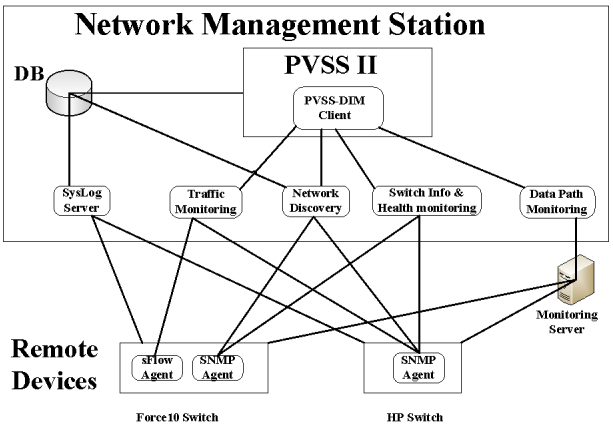

Figure 2: Architecture of the network monitoring system

NETWORK MONITORING SYSTEM

As part of the ECS, the network monitoring system provides heterogeneous interfaces to users based on the LHCb ECS framework.

Architecture of the Network Monitoring System

In the network monitoring system, PVSS II is used to monitor the status of the network devices and the throughput of switches. The system is mostly based on Simple Network Management Protocol (SNMP) [7] and sFlow [8] which are supported by all the network devices in LHCb. SNMP is an application layer protocol that facilitates the exchange of management information between managed network devices and the management station. SNMP enables network administrators to monitor and configure network devices. A SNMP driver is provided in PVSS II, but the performance is insufficient for a large system. To improve efficiency, a custom front-end process called SNMP collector has been developed. It is used to query interface counters and other information from network devices via SNMP, and transfer data to PVSS II via Distributed Information Management (DIM) [9] which is widely used in LHCb ECS. Because there are so many ports in the core switches, the SNMP query of interface counters takes a long time and occupies a lot CPU and memory resource. The sFlow collector has been developed to get the interface counters, while SNMP is still used for querying the other information of the core switches.

The SNMP collector logically consists of three primary modules: SNMP message transport module, SNMP message processing module, and DIM message exchange module. The SNMP message processing module deals with encoding and decoding of SNMP messages, and the SNMP message transport module is used for exchanging SNMP messages over network based on Berkeley sockets. Both the SNMP message transport module and message processing module are implemented based on the Net-SNMP [10] library. In the managed network devices, the properties of the managed object are collected in MIB (Management Information Base), which is accessible through an object identifier (OID).

The sFlow collector has a similar structure as the SNMP collector. SFlow is a sampling mechanism to capture traffic data in high speed networks. The sFlow agent is the implementation of the sampling mechanism based on hardware. There are two kinds of sFlow samples: flow samples and counter samples. Flow samples contain the header and some octets of the sampled packets based on a defined sampling rate. Counter samples contain the octet and packet counters for interfaces, and will be sent to the collector at a defined polling interval. The sFlow collector extracts the required counters from flow samples and sends to PVSS II via DIM.

In the network monitoring system, the SNMP collector and sFlow collector act as DIM servers, and PVSS II subscribes to the DIM services via PVSS-DIM bridge. In PVSS II, the information from switches are stored as a datapoint (DP) of a pre-defined datapoint type (DPT) which describes the data structure. A callback function is called to parse the raw data when the datapoint is updated. The state of the online network is represented by the FSM based on the LHCb ECS framework. In the FSM tree, the following are monitored: hardware status, traffic throughput status, uplinks between these switches, and the routing status of the data paths.

Network Discovery

The network topology discovery is based on the Link Layer Discovery Protocol (LLDP) [11], which is supported by all the network devices in LHCb online system. LLDP is a vendor-neutral Layer 2 protocol that allows a network device to advertise its identity and capabilities on the local network, and to learn about each other. The SNMP collector queries the LLDP MIB from the network devices via SNMP. With a seed device as a starting point, the LLDP neighbors will be discovered, and then the neighbors of those neighbors, and so on until all the devices have been discovered in the network. Based on the LLDP information, the map of the network topology can be learned accurately. The status of uplinks are communicated to PVSS II via DIM, an alarm will be issued if any uplink is down.

To discover the attached nodes in the network, the SNMP collector queries the Address Resolution Protocol (ARP) entries and the Media Access Control (MAC) addresses learned by switches. MAC address is also known as Ethernet hardware address which is unique for each network device. ARP entry consists of the IP address and the corresponding MAC address. With all the information of uplink, ARP and MAC addresses, the nodes can be discovered, including the IP address, MAC address, and the switch port to which it is currently attached.

All the topology information is stored in the database, which is used by PVSS II.

Traffic Monitoring

For the HP switches, the SNMP collector queries the interface traffic counters (ifInOctets, ifInPackets, ifOutOctets, ifOutPackets), and the interface errors counters (ifInError, ifOutDiscard). The interface traffic counters are used to calculate the bandwidth utilization for each interface along with the device uptime. The interface errors counters indicate the error of each port in the switch. In order to get an accurate traffic measurement and not to consume too much CPU and memory resource of switches, the query interval is set to 60 seconds for all the devices. For the core switches, the polling interval of sFlow is set to 60 seconds as well. sFlow counters samples provide the same counters as SNMP, and the sFlow collector provides the same interface to PVSS II as the SNMP collector.

Based on the trending tool of the JCOP framework, a panel has been developed to display the histograms of the instant input bandwidth utilization, output utilization, and the errors for each ports, which allows to see the overview of all ports in a switch. Another panel has been developed to display the traffic trending (including port utilization, packet rate, error) of the interface. All the traffic information is archived in an Oracle database. The system can generate the report for a given period by query the archived data. This can be used to analyze the traffic model and optimize the network.

Data Path Monitoring

The data path monitoring is specific to the data network, a tool has been developed to monitor the data paths. From the network's point of view, there are three stages for the LHCb DAQ: event data from readout boards to the HLT CPU farm, selected event data from the HLT farm to the LHCb online storage, raw data files from the LHCb online storage to CERN CASTOR. In all these stages, the data transfers are working at Layer 3. This tool sends Internet Control Message Protocol (ICMP) echo request packets to ordinary computer nodes and ARP requests to readout boards which don't support ICMP, then listens to response packets. At the end of the scanning for each stage, the result is summarized and published to PVSS II via DIM.

Syslog Server

A syslog [12] server is setup to receive the syslog messages from the network devices. Syslog allows a machine to send the event notification messages across IP networks to the event message collector. The syslog packets are sent in clear text, and have three distinct parts: priority, header and message. The priority part represents both the facility and severity of the message. A facility in syslog is a class of messages, the facility LOG_NEWS is assigned to the network devices in the LHCb online system.

When network devices run into problems, error messages will be generated and sent to the syslog server as configured in the network device, these include system errors, parameters reach threshold, hardware failure and so on. The syslog server logs all the received messages into a text file. For the message with a higher priority above warning, the syslog server stores the message in the database, and notices PVSS II via DIM.

Health Monitoring

In this part, some parameters indicate the health condition are monitored, these parameter includes the temperatures inside switches, fan speed, power supply. Besides, the utilization of CPU and memory are monitored as well. When the parameter reach the defined threshold, an alarm will be generated.

CONCLUSION

This paper describes the features and implementation of the network monitoring system based on SCADA system PVSS II and the JCOP framework. The system works in an efficient way with the custom front-end processes to query the SNMP MIB and collect sFlow counters samples from the network devices. With the tools provided by the JCOP framework, the FSM represents the states in network domains, and provides a common look and feel for the operators. The online network is monitored at varied levels. However, there are still some improvements to be done. Some interactive behaviors e.g. reset, should be implemented in the FSM for the experiment operators. The autogenerated daily report and long-term analysis report should be implemented.

REFERENCES

[1] LHCb Collaboration,"LHCb Technical Proposal", 1998.

[2] LHCb Collaboration, "LHCb Online Technical Design Report", 2001.

[3] C. Gaspar and B. Franekand R. Jacobsson and B. Jost and N. Neufeld and et al, "An integrated experiment control system, architecture, and benefits: the LHCb approach", IEEE Transactions on Nuclear Science, vol. 51, p. 513 2004.

[4] S. Schmeling at al, "Controls Framework for LHC experiments", IEEE-NPSS Real Time Conference, Montreal, Canada, 2003.

[5] "PVSS II", `http://www.pvss.com/`.

[6] C. Gaspar, "PVSS & SMI++ Tools for the Automation of large distributed control systems", ICALEPCS 2005.

[7] W. Stallings, "SNMP, SNMPv2, SNMPv3, and RMON 1 and 2 (Third Edition)", Addison-Wesley Professional, 1999.

[8] P. Phaal and S. Panchen and N. McKee, "InMon corporation's sFlow: a method for monitoring traffic in switched and routed networks (RFC 3176)", 2001.

[9] C. Gaspar, "DIM: Distributed Information Management", `http://dim.web.cern.ch/dim`.

[10] "Net-SNMP", `http://www.net-snmp.org`.

[11] "Link Layer Discovery Protocol", IEEE 802.3ab, 2005.

[12] C. Lonvick, "The BSD Syslog Protocol (RFC 3164)", 2001.