FERMILAB-CONF-07-114-CD-E

The DZERO Level 3 DAQ System: Operation and Upgrades

Aran Garcia-Bellido*, Tulika Bose†, Gustaaf Brooijmans‡, Doug Chapin†, David Cutts†, Stuart Fuess§, Thomas Gadfort*, Andrew Haas‡, William Lee§, Ron Rechenmacher§, Scott Snyder¶, Gordon Watts* and Yunhe Xie†

*Department of Physics, University of Washington, Seattle, WA 98195, USA
†Department of Physics, Brown University, Providence, RI 02912, USA
‡Department of Physics, Columbia University, New York, NY 10027, USA
§FNAL, Batavia, IL 60510, USA
¶Physics Department, Brookhaven National Laboratory, Upton, NY 11973, USA

***Abstract*— The DØ Level 3 data acquisition system for Run II of the Tevatron has been reliably operating since May 2002. Designed to handle average event sizes of 250 kilobytes at a rate of 1 kHz, the system has been upgraded to be able to process more events, doubling its typical output rate from 50 Hz to 100 Hz, while coping with higher event sizes at the beginning of high luminosity collider stores. The system routes and transfers event fragments from 63 VME crates to any of approximately 320 processing nodes. The addition of more farm nodes, the performance of the components, and the running experience are described here.**

I. Introduction

The DØ trigger and data acquisition systems are designed to operate at the high instantaneous luminosities achieved in Run II of the Tevatron proton-antiproton collider. The first level of triggering decisions reduces the collision rate of 1.7 MHz to an output of around 2 kHz, based on partial information from the subdetectors. At the next trigger level, having more refined information, the rate is further reduced to around 1 kHz. These first two levels of the trigger use primarily hardware and firmware to execute the decisions, although the second level also uses software for its event reconstruction and triggering decisions. The last level of the trigger, Level 3, is based on software running in a farm of more than 300 computers. The Level 3 data acquisition system (L3DAQ) [1], after a Level 2 accept decision is made, is designed to transfer the event fragments from 63 VME readout crates to the processing nodes of the Level 3 farm. The total event size can reach more than 300 kilobytes, with each VME crate containing 1-20 kB distributed among VME modules. Those event fragments must be routed to one of the processing farm nodes where they will be concatenated and a Level 3 decision will be made with access for the first time to the full detector readout information. The final output rate is around 100 Hz of events saved for offline analysis, or 30 MB/s.

The full L3DAQ system is built with commodity hardware running the Linux operating system and based on Ethernet communication. The system is built around a single Cisco 6509 Ethernet switch [2]. Each VME crate contains a single-board computer (SBC) that reads out the VME modules and sends the data to one of the farm nodes as specified by routing instructions obtained from a dedicated SBC called Routing Master (RM). Each Level 3 farm node sends information to the RM about its load and the RM makes a decision of which event to send to which node. Once all the fragments arrive at node an Event Builder (EVB) process collates the fragments into complete events and these are placed in shared memory buffers for processing by several Level 3 filter processes. A schematic view the components is shown in Fig. 1 and the operation and data flow is shown in Fig. 2.

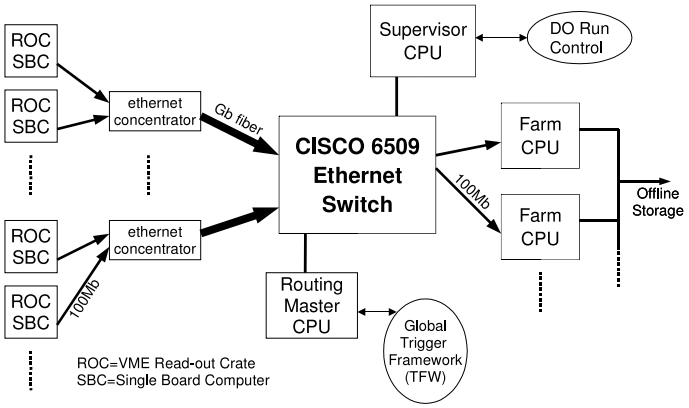

Fig. 1. The physical network configuration of the L3DAQ system

A detailed description of the system can be found in Ref. [1], here we will describe recent updates, the current performance of the system and some future upgrades.

II. Running Conditions

The peak instantaneous luminosity of the Tevatron has steadily increased since 2003 from less than 100 10^{30} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ to almost 300 10^{30} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$, and is planned to increase still a bit more in the coming future. The event size has grown accordingly, due to the higher occupancy of the detectors (because of busier events with increased particle multiplicity in them). At the highest current luminosities the average event size has peaked at 350 kB, decreasing to around

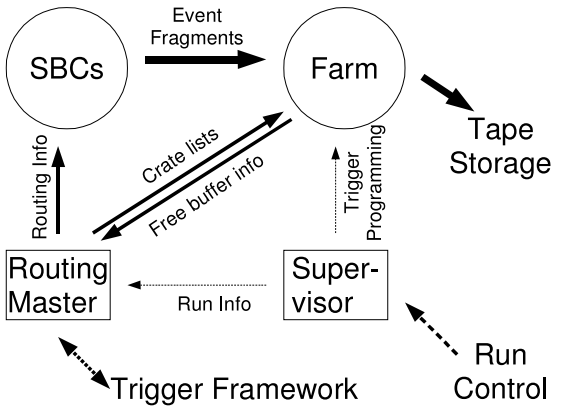

Fig. 2. Schematic illustration of the information and data flow through the L3DAQ system.

250 kB for the lowest luminosities. The L2 accept rate has rarely gone beyond 1 kHz, its design value. Before 2006 the L3 output rate was limited to 50 Hz, while we now routinely operate at 100 Hz or more at the beginning of the store. This has been made possible mainly by the addition of more farm nodes (from 82 in 2004 to the current 328) and other adjustments in the system.

III. Single Board Computers Upgrades

We currently have three different configurations of the Ethernet ports in the running SBCs. Fifty SBCs, whose data load is smaller than 12 kB per event fragment, remain with a single 100-Mb Ethernet interface active. At 1 kHz and 12 kB per fragment, the saturation limit of a single 100-Mb Ethernet connection, around 10 MB/s, would be reached. Thirteen crates, those with a data load above 12 kb, operate with two 100-Mb Ethernet interfaces, where the farm nodes establish two TCP/IP socket connections and the SBC alternates between the two connections. The transfer limit for two 100-Mb Ethernet interfaces is reached at 25 MB/s, or 25 kB at 1 kHz. Three crates, with fragment sizes beyond or around 20 kB use Gb connections, with direct connections to the Cisco switch. These SBCs have been equipped with a Gb card.

Figure 3 shows the performance of dual 100 Mb Ethernet SBCs as a function of the fragment size, in a test environment. Figure 4 shows the yearly average and peak data transfer rates of an SBC with a Gb Ethernet connection.

IV. Routing Master Upgrades

The Single Board Computer that houses the Routing Master process was upgraded from the usual VMIVME 7750 with a Pentium III at 933 MHz to a VMIVME 7805 [3] with a faster Pentium 4M processor at 1.7 GHz. The RM decisions are now always made in under 1 ms, and in a test environment we could bring the CPU to its maximum only at 1.4 kHz input rate to L3DAQ, which is well beyond the maximum design conditions of 1 kHz.

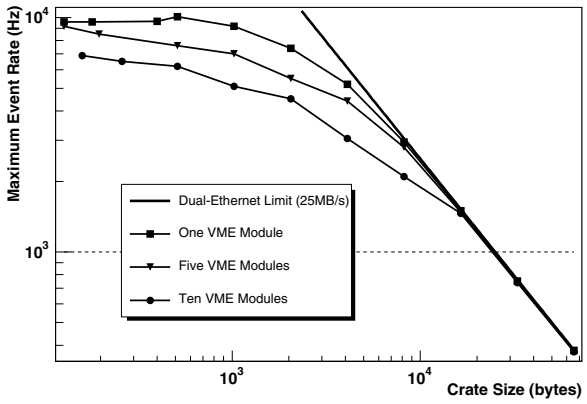

Fig. 3. Dual 100 Mb Ethernet port use for one SBC in test conditions. The maximum event rate as a function of the data fragment size being read by that SBC. It can be seen that in normal conditions (below 1 kHz event rate), the dual Ethernet limit of 25 MB/s is reached when the fragment size is around 20 kB or larger.

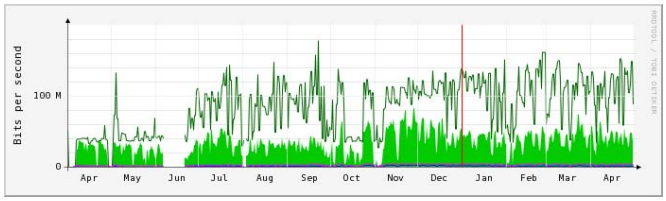

Fig. 4. Yearly traffic report from an SBC with Gb Ethernet, as reported by the Cisco switch. The green filled histogram is the incoming traffic from the SBC and the dark green line is the peak incoming traffic, in Mb/s.

V. Upgrades to the L3 Farm

The L3 farm has grown from 82 computers in 2004 to the current 328 in order to accommodate the changing physics conditions and to efficiently take data at the highest instantaneous luminosities. The L3 software that reconstructs and triggers the event has evolved as well with time since it takes longer to reconstruct an event the higher the instantaneous luminosity.

The L3 output rate, controlled by "trigger suites" that scale which types of events are accepted as a function of the luminosity, has increased from around 50 Hz to 100 Hz. The limit being here the cost of tapes to store the recorded events, more than any hardware limitations.

Table I shows the four different purchases of nodes and their specifications currently in use at DØ. In all new purchases the Fermilab Computing Division and DØ Online Group have assured the quality of the computers, made the installation in racks, tested and burnt-in the machines, and are responsible for the maintenance and monitoring of the machines.

The number of L3 filter processes running in each type of node has been optimized to increase the capacity of the farm. As shown, in Tab. I, dual processor single core hyperthreaded chips (with four "effective" processors) are only allowed to run three filter processes. The net gain from using two to using three processes in these machines is around 20% more events

processed. If four processes were to be used, the gain would remain the same as with three, but the memory would be more heavily taxed. Dual processor dual core machines (with four real processors) can run four filter processes naturally.

The performance of the different node types in the farm as a function of instantaneous luminosity is presented in Figs. 5 and 6, for a recent store (see Ref. [4] for a description of the L3DAQ monitoring) . It can be seen here the CPU usage by the L3 filters and the average filter processing time per event. The CPU usage scales similarly with luminosity for the four different types of nodes. But the average processing time per event shows differences between the four types of nodes, according to how many filter processes are running, the type of chips and the memory bandwidth. The scaling of the dual core chips with luminosity is less steep than for single core hyperthreaded chips. The comparison is not favorable to the latter, since they run more filter processes than they have real processors.

The maximum number of event fragments in flight to any farm node at any moment was initially set to three, to avoid filling up the TCP send buffers on the SBC. This number was calculated assuming a very large crate size of 256 kB, many times over the normal physics running conditions, and the old number of farm nodes. This buffering scheme requires that the number of advertised buffers multiplied by the crate size multiplied by the number of farm node connections in the SBC is less than the available memory in the SBC. The new number of maximum event fragments in flight to any node has been set to six, this is how many buffers each node can tell the RM it has available. The change was implemented to avoid the nodes backing up during calibration runs with a limited number of nodes. In normal data taking with the full farm available, usually only one buffer is used per node and the limit of six is never reached. With around 100 MB available memory, six advertised buffers and a maximum of 2×328 farm node connections on those SBCs with dual Ethernet in use, the maximum crate size would be around 25 kB, which for those SBCs with dual Ethernet is never reached in normal data taking.

VI. Performance of the Cisco 6509 Switch

The Cisco switch has 9 blades and is currently fully used with 7 blades that support the connection to 48 L3 farm nodes each. Another blade receives the Gb connections from the Ethernet concentrators, the three individual Gb SBCs and connects to the online network and controller processes. The final blade is used for the switch supervisor process.

Each of the 7 slots connected to 48 farm nodes have a maximum of 112 MB of shared memory. The current configuration, with a maximum of six advertised buffers on the farm nodes and maximum event size of 350 kB gives around 100 MB of data in transit through each blade. Only if the average event size exceeds 350 kB the switch shared memory limit will be reached and packets may be dropped. In normal operation, we operate well below the 112 MB limit.

VII. Running Experience

The SBCs have proved to be extremely reliable hardware. Very few repairs have been needed since 2002, maybe one or two per year initially, but we have had no hardware problems in the last two years. The onboard compact flash disk of 128 MB is very slow and is only used to store configuration information. The SBCs are network booted and use the RAM disk to load the OS and running processes. NFS mounts are used to access the software running on the SBC and also to store log files. For monitoring and maintenance, it would be desirable to have a console server, but we have managed with a portable keyboard and monitor.

The farm nodes have had frequent hardware problems, usually on the hard drives or CPU fans, but this strongly depends on the quality of the components. Some batches really come with an outstanding durability. In others, a few machines have problems from the beginning and are difficult to salvage. But overall, we usually run with less than 10 unused farm nodes. We deal with minor problems, like hangs or disconnections, usually recovered by a reboot, on a few nodes per week. More serious problems, that require a hardware maintenance call, occur at the few nodes per month.

One area that has become increasingly problematic is the versioning of the L3 filter software running on the farm. The size of the executables, scripts, and configuration files is around 340 MB and they are distributed to all nodes via rsync and rgang [5], such that if a small modification is implemented it can be easily shared with all machines. But the first installation of a new version of the software now takes a very long time, having to chain copy 340 MB to more than 320 machines. New installations currently happen around three times a year and the version that all nodes are running is set manually through shell scripts. Tools like BitTorrent [6] together with Cfengine [7] or some peer-to-peer management software would ease the versioning and control of the software running in the farm, by having a "master" node that the others can synchronize to immediately after they boot.

Since problems occur frequently with farm nodes, it is essential that the rest of the software can dynamically adapt to the loss of one or or more farm nodes at any time, and specially during data taking, without interrupting the data flow. The RM, the SBCs and the nodes must know at all times which of them are active and responding. If one farm node is unresponsive, the supervisor process detects it and notifies the RM not to route events to that node. Farm nodes as well send TCP keepalive pings to the RM and the SBCs.

VIII. Conclusion

The L3DAQ system has undergone a few upgrades since it was commissioned in the spring of 2002. Due to the original design, these changes have been easily implemented to accommodate the higher instantaneous luminosities of the Tevatron and the L3 trigger physics requirements. The number of farm nodes has quadrupled since 2004, and the input rate has reached close to 350 MB/s from the original design of 250 MB/s. The L3DAQ system has transitioned smoothly

TABLE I

THE COMPOSITION OF THE L3 FARM OF COMPUTERS

Name	# of boxes	CPU	L3 filter processes
ASA	32	Double processor Pentium Xeon 2.8GHz	hyperthreaded, running three filters
KOI	128	Double processor Pentium Xeon 2.8GHz	hyperthreaded, running three filters
CAB	48	Double processor double core AMD Opteron 1.8 GHz	running four filters
ACE	120	Double processor double core Pentium Xeon 2.3 GHz	running four filters

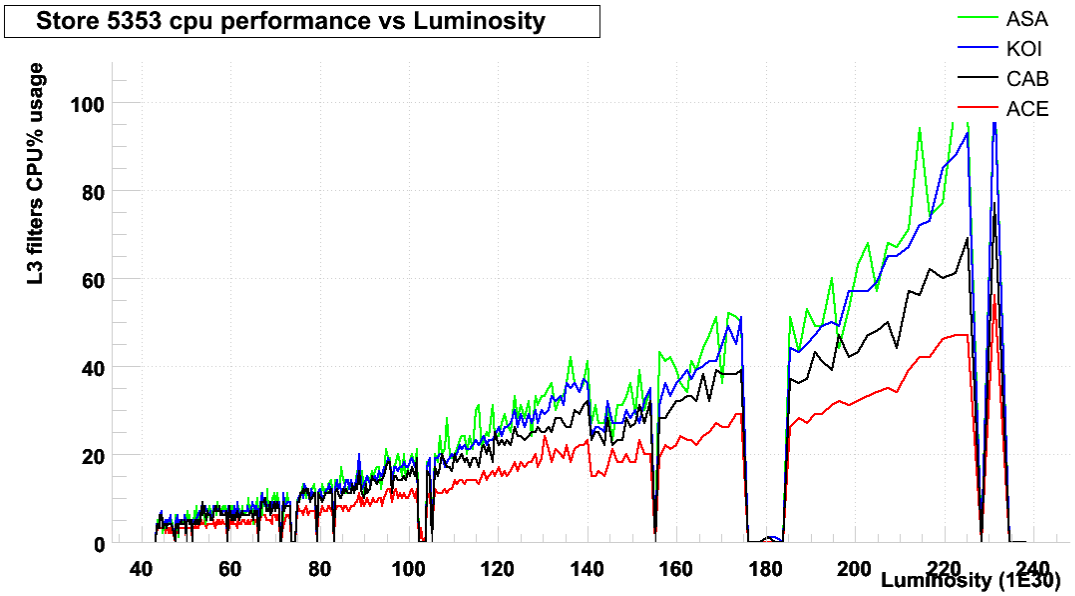

Fig. 5. L3 farm nodes average CPU performance during a store, as a function of the instantaneous luminosity, for the different types of nodes in the farm.

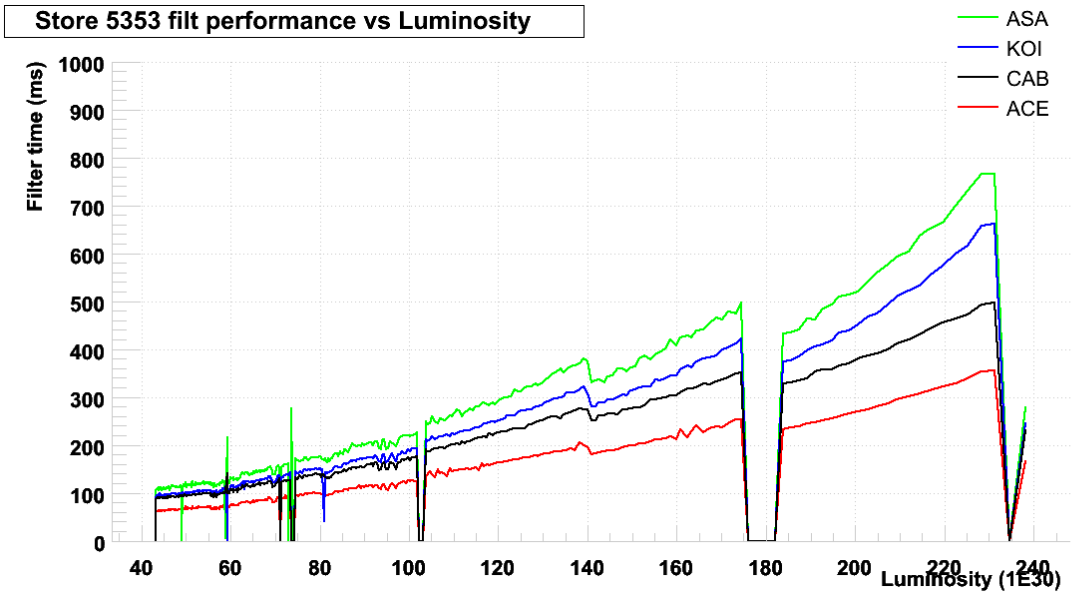

Fig. 6. The average time it takes a L3 filter to process one event as a function of the instantaneous luminosity during a store, averaged for the different types of nodes in the farm.

through all these changes and continues to serve the physics needs of DØ.

ACKNOWLEDGMENT

The L3 farm of computers is maintained by the DØ Online Group and the Fermilab Computing Division: they do an extraordinary job from installation to repairs. The authors wish to acknowledge their dedication and support, and thank them for their contribution to the high efficiency of data taking at DØ.

REFERENCES

[1] R. D. Angstadt *et al.*, "The DZERO level 3 data acquisition system," IEEE Trans. Nucl. Sci. **51**, 445 (2004).
[2] `http://www.cisco.com`
[3] `http://www.gefanucembedded.com`
[4] A. Haas *et al.*, "D0 online monitoring and automatic DAQ recovery," *In the Proceedings of 2003 Conference for Computing in High-Energy and Nuclear Physics (CHEP 03), La Jolla, California, 24-28 Mar 2003, pp THGT004* [arXiv:physics/0306195].
[5] `http://fermitools.fnal.gov`
[6] `http://www.bittorrent.com`
[7] `http://www.cfengine.org`